SUPPLEMENTO
AL VOLUME XIII, SERIE X, DEL

NUOVO CIMENTO

A CURA DELLA SOCIETÀ ITALIANA DI FISICA

1959 3º Trimestre N. 2

RENDICONTI

DEL

VII CORSO CHE NELLA VILLA MONASTERO A VARENNA

DAL 7 AL 19 LUGLIO 1958

FU TENUTO A CURA

DELLA SCUOLA INTERNAZIONALE DI FISICA

DELLA SOCIETÀ ITALIANA DI FISICA

SULLA

TEORIA DELLA INFORMAZIONE

INDICE

INTRODUCTION

LECTURES

SEMINARS

La direzione del Corso porge vivi ringraziamenti ai dott. Valentino Braitenberg *e* Nello Onesto *di Napoli e* Anna Cuzzer *di Roma per l'aiuto dato prima nell'organizzazione del Corso, poi nello svolgimento di esso e infine nella compilazione di questo fascicolo, per il quale particolari ringraziamenti vanno anche al dott.* Sergio Ratti *di Milano.*

INTRODUCTION

Parole inaugurali

DI

Giovanni Polvani

Presidente della Società Italiana di Fisica

Rinnoviamo oggi a quindici giorni di distanza questa breve cerimonia inaugurale per questo II Corso, 1958 — VII della serie iniziata nel 1953 —, il quale, dopo quello chiusosi sabato passato relativamente alla Fisica del plasma, viene questo anno tenuto, a cura della Scuola Internazionale di Fisica della nostra Società Italiana di Fisica, sulla Teoria dell'Informazione, disciplina di grande attualità e novità e certamente di grande possibilità di applicazione ai più svariati campi sia della conoscenza pura sia della tecnologia, comprendendo in questo termine anche i mezzi di scambio, di critica e di produzione della stessa conoscenza.

E, come le altre volte, estremamente gradito mi è porgere, anche a nome della Società, il saluto più cordiale a tutti i presenti, specie a S. E. l'on. Scaglia, Sottosegretario alla Pubblica Istruzione, intervenuto in rappresentanza del Ministro prof. Moro; e porgere anche a tutti i partecipanti al Corso il benvenuto alla nostra Scuola e a Varenna.

I quali partecipanti non sarebbe del tutto corretto distinguere, come vorrebbe la consueta organizzazione dei Corsi della nostra Scuola, nelle categorie di docenti, allievi e uditori.

La ragione è che questa nuova disciplina, la Teoria della Informazione, se proprio disciplina autonoma si può chiamare —, è ancora in via di formazione. È quindi opportuno che le « lezioni » vere e proprie possano talora cedere il passo, se necessario, alle discussioni, all'esposizione di risultati appena conseguiti o di orientamenti intuiti in ricerche originali in corso...: in altre parole occorre che la « scuola » si trasformi in « convegno » (in « simposio », oggi si direbbe) o addirittura in pura e semplice « conversazione scientifica ». I partecipanti potrebbero allora esser considerati sotto l'unico aspetto di « studiosi »; e, volendo mantenere il termine di « docente », « docenti » possiamo allora chiamare quelli che si sobbarcheranno alla fatica di tenere alcune lezioni, svolgere conferenze e seminari, portare il contributo critico della propria scienza e perizia nel vivo della discussione, nei momenti di dubbio, di esitazione ecc.; e gli altri chiamare « allievi » o, meglio, « studenti » solo perchè vogliono, soprattutto, studiare, apprendere, penetrare questa nuova disciplina che è la Teoria dell'Informazione.

Accettata questa divisione, lasciate che, come alle altre inaugurazioni, io li presenti reciprocamente questi partecipanti al Corso, questi « ospiti » — chè questo è il termine pieno —, questi ospiti graditissimi; e lasciate anche che alla regola generale di nominarli in ordine alfabetico faccia un'eccezione, certo da tutti approvata e a tutti gradita, nominando per primo chi tra i docenti sarebbe ultimo... alfabeticamente.

Docenti: NORBERT WIENER di Cambridge (U.S.A.); Y. BAR-HILLEL di Jerusalem, V. BRAITENBERG di Napoli, R. BUSA di Gallarate, E. R. CAIANIELLO di Napoli, W. DAVENPORT di Lexington (U.S.A.), R. M. FANO di Cambridge (U.S.A.), A. FEINSTEIN di Stanford (U.S.A.), D. GABOR di Londra, P. E. GREEN di Lexington, M. HALLE di Cambridge (U.S.A.), B. HASSENSTEIN di Tubinga, H. HAUS di Cambridge (U.S.A.), D. A. HUFFMANN di Cambridge (U.S.A.), Y. W. LEE di Cambridge (U.S.A.), L. LÖFGREN di Stoccolma, B. MCMILLAN di Murray Hill (U.S.A.), B. MANDELBROT di Paris, G. MORUZZI di Pisa, E. NEWMAN di Cambridge (U.S.A.), W. REICHARDT di Tubinga, R. RIGHI di Roma, N. ROCHESTER di Yorktown Heights (U.S.A.), W. ROSENBLITH di Cambridge (U.S.A.), J. SCHOUTEN di Eindhoven, M. SCHUTZENBERGER di Parigi, D. SLEPIAN di Murray Hill (U.S.A.), F. STUMPERS di Eindhoven, S. WATANABE di Ossining (U.S.A.).

A questi che ho nominato il ringraziamento più vivo per la loro collaborazione, specie all'amico prof. EDUARDO R. CAIANIELLO che, sobbarcatosi già alla non piccola fatica dell'organizzazione scientifica del Corso, si sobbarca ora a quella della direzione: a lui esprimo, anche a nome della Società tutta, un ringraziamento particolarmente caloroso ed affettuoso per tutto quello che ha fatto e farà per l'attuazione del Corso.

Studenti: C. BOHM di Roma, F. BRESSON di Parigi, J. C. BRIANNE di Lilla, M. CECCARELLI di Bologna, G. COLOMBO di Milano, A. CUZZER di Roma, A. FIORENTINI di Arcetri, J. HILL di Redding, F. DE JAEGER di Eindhoven, F. LAURIA di Napoli, A. LEPSCHY di Roma, L. LUNELLI di Milano, L. MONTANET di Genève, A. L. NAGEL di Eindhoven, H. OHZU di Vienna, N. ONESTO di Napoli, F. PANDARESE di Napoli, F. PIERANTONI di Bologna, G. QUAZZA di Milano, A. RUBERTI di Roma, C. SCHAERF di Roma, W. F. SCHALKWJIK di Eindhoven, P. SCHNUPP di Gottinga, C. VAN SCOONEVELD dell'Aja, V. SOMENZI di Roma, A. J. STAM dell'Aja, L. STIBE di Cambridge

(U.S.A.), D. Variu di Tubinga, V. Vittorelli di Palermo, J. Witt di Monaco, R. Woodcock di Hampton.

A tutti l'augurio più vivo di trarre il miglior profitto da questo Corso-Convegno.

Non voglio infine perdere l'occasione datami da questa riunione inaugurale per rinnovare le espressioni di più viva gratitudine verso tutti coloro che col loro aiuto finanziario hanno messo la Scuola in condizioni di potere compiere quest'anno questo sforzo veramente notevole di organizzare e svolgere quattro corsi. Già nominai i nostri sovvenzionatori nel mio breve discorso d'inaugurazione del I Corso 1958 chiusosi ieri l'altro, sabato; ed è forse superfluo che io li ricordi nominativamente uno per uno; ma al Massachusetts Institute of Technology di Cambridge degli Stati Uniti, che si è sobbarcato le spese di viaggio dei molti docenti a questo Corso che provengono dall'America, e all'Università di Napoli che ha efficacemente aiutato a sostenere le spese di preparazione organizzativa del Corso, desidero porgere, anche a nome della Società, un particolare ringraziamento.

E ormai chiudo questo mio discorso. Ieri sera parlando con alcuni amici ho sentito esprimere la loro gradita sorpresa per la varietà delle nazioni che in ognuno di questi corsi internazionali sono rappresentati: diciotto in quello passato, undici in questo. È uno dei meriti della scienza in generale e della Fisica in particolare il saper gettare ponti ben saldi al disopra delle così tante separazioni che col nome di « confini » dividono gli uomini dagli uomini. E non potremmo allora fondatamente sperare (o invece è proprio follia sperare) che ciò possa essere domani la via per la quale finalmente tutti gli uomini troveranno una base comune di sicura e cordiale convivenza? Se a questo — come è auspicabile — e per questa via un giorno si potrà arrivare, anche la nostra Scuola Internazionale di Fisica, anche Varenna avranno il loro piccolo merito.

Con questi sentimenti dichiaro aperto il II Corso di Fisica 1958 — VII dall'inizio della Scuola Internazionale di Fisica della nostra Società — relativo alla Teoria della Informazione.

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3° Trimestre

Prolusione

DI

EDUARDO R. CAIANIELLO

Direttore del Corso

The seventh Course of the International Summer School of Physics at Varenna is the first ever held here on the Theory of Information. Let me welcome first of all the lecturers and guests who have braved oceans and mountain passes to be with us here for the coming two weeks.

We have a few Italian lecturers and many Italian students here, and I welcome them as an encouraging sign that one aim of this school is being achieved, namely to further the growing interest for this new field among Italian Universities.

We hope that this will not turn out all too strictly speaking to be a school composed only of transmitters and receivers of knowledge. There will be ample opportunity for argument and, we hope, for the building of new bridges among the representatives of various branches of cybernetics. To achieve this end, the presence of Professor NORBERT WIENER, to whom we owe more than just the name of this science, is especially auspicious. And we are glad that you have by general acclamation elected him as Permanent Chairman of our meetings.

Most meetings are much more work to prepare than this one has been, which could avail itself of the smooth-running organization of the Summer School of Physics at Varenna of the Società Italiana di Fisica. Our special thanks go to Professor GIOVANNI POLVANI, President of the Italian Physical Society, whose untiring efforts have created this physicist's paradise.

Last, certainly not least, let us thank the Ministero Italiano della Pubblica Istruzione, the Società Italiana di Fisica, and the Massachusetts Institute of Technology, without whose magnificent sponsorship this meeting would not have been possible.

* * *

La Cibernetica — di cui la Teoria della Informazione costituisce il nucleo matematico — ha solo di recente acquisito fisionomia e nome di scienza autonoma. Essa ha come oggetto l'indagine delle relazioni esistenti tra le varie parti di un organismo, prescindendo dai particolari costruttivi di ciascun elemento, che viene caratterizzato esclusivamente mediante la sua funzione.

Se chiamiamo « anatomia » di un organismo — sia esso una macchina o una società industriale o un essere vivente — lo studio della sua costituzione particolareggiata, potremo allora denominarne « fisiologia » lo studio cibernetico; o, con altra analogia, possiamo paragonare il passaggio dallo studio strutturale a quello cibernetico, alla transizione dall'Aritmetica all'Algebra.

Come appunto succede allorchè nelle relazioni matematiche, i numeri particolari vengono sostituiti con simboli algebrici, si scoprono proprietà generali prima insospettate; così organismi diversissimi rivelano di possedere identico funzionamento, e metodi matematici per lo studio diretto di relazioni funzionali vengono via via opportunamente elaborati.

Il funzionamento di un organismo complesso richiede che ciascun costituente di esso abbia conoscenza del comportamento degli altri costituenti; le comunicazioni tra le varie parti di un sistema sono quindi di fondamentale interesse, e il loro studio — Scienza delle Comunicazioni — viene volta a volta riguardato, a seconda dei punti di vista, come parte essenziale della Cibernetica, o come addirittura coincidente con essa.

Ciò che viene comunicato è una informazione; la precisazione quantitativa di questo concetto è il punto di partenza della Teoria dell'Informazione, che elabora l'apparato matematico necessario agli studi cibernetici.

Si comprende dunque come il campo aperto a tali indagini sia vastissimo. Scegliendo alcuni esempi a caso, questioni quali la più efficiente utilizzazione di una linea di trasmissione, l'invenzione di calcolatrici elettroniche o di macchine atte a dimostrare teoremi di Logica, la traduzione meccanica dei linguaggi, l'analisi quantitativa dei fenomeni nervosi, possono dare un'idea della varietà pressocchè illimitata di temi che la Cibernetica si propone di trattare con metodo unitario.

Mentre tutte le altre scienze si vanno differenziando e specializzando sempre più e creano linguaggi mutuamente incomprensibili, la Cibernetica si presenta come un tentativo di sintesi, un ponte gettato tra molti rami del sapere, una disciplina che vuole intendere, collegare e coordinare teorie e fatti propri delle altre o di altre discipline, mediante la scoperta di funzioni comuni in oggetti di natura diversissima. A seconda dell'interesse del ricercatore, essa può intendersi come un raffinamento di principi di tecnica elettronica, o, via via innalzandosi, come un nuovo umanesimo scientifico che studia il comportamento di collettività umane, o la genesi e la simulazione del pensiero, su basi strettamente quantitative.

Immenso sviluppo hanno queste ricerche nei paesi più avanzati scientificamente, Stati Uniti e Russia; governi e industrie le appoggiano in tutti i modi, ed è facile prevedere che, tra non molto, attraverso di esse, si giungerà ad una rivoluzione industriale sconvolgente almeno quanto la prima, che sostituì il lavoro della macchina al lavoro muscolare: il lavoro intellettuale « di tipo non creativo » sarà svolto dalle nuove macchine cibernetiche (come già sta avvenendo in diversi campi).

L'Italia è stata finora totalmente assente da questa attività di ricerche e applicazioni: circostanza questa particolarmente dolorosa, in quanto molte di queste indagini si possono compiere con dispendio minimo di denaro.

Il Corso, che ho l'onore di dirigere, è stato concepito e organizzato nella speranza che esso possa segnare l'inizio, anche da noi, di una seria attività di studio e di ricerca in questo campo, e di un'attiva collaborazione in ambito sia nazionale, sia internazionale.

Esso ha perciò una fisionomia particolare: come accennava or ora il Presidente della Società Italiana di Fisica nel suo discorso di apertura, invece della consueta distinzione in docenti, allievi e uditori, e del consueto rapporto numerico tra di essi, noi abbiamo in questo Corso solo due gruppi, egualmente numerosi (di circa trenta persone l'uno): il primo è costituito in gran parte da scienziati di chiara fama internazionale, convenuti qui da molte parti del mondo; il secondo da partecipanti, la maggior parte dei quali possono dirsi studenti solo perchè desiderosi di apprendere dai primi i fondamenti di una nuova scienza.

È questo, dunque, in parte un Convegno, in parte una Scuola: le ore pomeridiane saranno dedicate in prevalenza a comunicazioni originali, a discussioni, a seminari; quelle mattutine a lezioni di carattere istituzionale. Questo Corso-Convegno è anche, nell'ambito degli studi cibernetici, la prima iniziativa così impostata ed attuata; e la qualità degli studiosi che hanno voluto onorarci della loro presenza e la splendida tradizione che ormai si ricollega al nome della Scuola estiva di Varenna, sono certo auspicio del suo felice successo.

SOCIETÀ ITALIANA DI FISICA

SCUOLA INTERNAZIONALE DI FISICA

7° CORSO ESTIVO - VARENNA SUL LAGO DI COMO - VILLA MONASTERO - 7-19 Luglio 1958

A Descriptive Introduction to the Statistical Theory of Communication.

B. McMillan

Systems Engineering Division, Bell Telephone Lab. - Murray Hill, N.J.

Introduction.

As an introduction to the subject of this Course I propose to attempt a classification of some of the problems considered in the theory of communication and to relate these problems to each other, and to the problems considered in conventional statistics.

Three domains of discourse will be considered, each briefly:

1) Statistical inference,

2) Communication as related to measurement and control,

3) Communication as a service (*e.g.* telegraphy),

The modern approach to statistical inference began to develop in, roughly, 1900. An important stimulus to its growth has been the need of experimenters in agriculture for tools to handle highly variable data. It suffices for this introduction to divide the problems considered under this heading into two classes:

A: Testing hypothesis,

B: Estimation of parameters.

In fact, these divisions are neither exhaustive nor mutually exclusive.

The typical problem of testing (A, above) is illustrated by the following agricultural experiment: several plots of ground, as nearly alike as possible, are selected and prepared for planting. Half of them are treated with fertilizer, the other half not treated. They are then planted and cultivated alike. When the crops are harvested, the yields from each plot are determined. Given the several yields $y_1, y_2, ..., y_n$ from the treated plots, and the yields $z_1, z_2, ..., z_n$ from the untreated plots, the statistician is then asked to test, *i.e.* to accept or reject, the so-called « null hypothesis »: the hypothesis that the fertilizer has no effect. His method, of course, is to reject the hypothesis if the y's appear « too different » from the z's; the critical degree of difference is determined by the internal variations among the y's, and among the z's.

Underlying the statistical treatment of this experiment is a mathematical model which can be diagrammed thus:

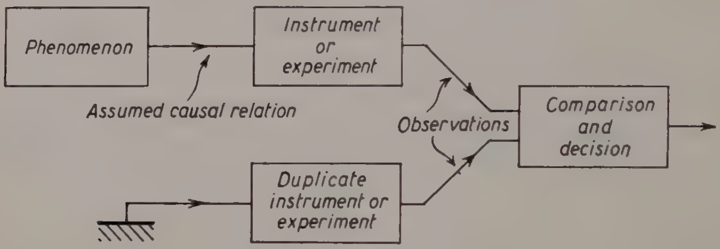


Fig. 1.

This diagram symbolizes the fact that the phenomenon of interest is observed through an intervening instrument or experiment with which it is presumed to be causally related. In general the instrument has two important defects which prevent the observations it develops from representing exactly the phenomenon being studied. In the first instance, it is typical that the instrument distorts, *i.e.*, presents an image of the phenomenon under study which is not an exact copy and even may be so distorted that some features of the original phenomenon are destroyed. That is, it may be that some features of the original cannot be recovered or determined from even a perfect knowledge of the image. This defect does not directly interest the statistician, though it may be of great importance to the experimenter, or to one studying the theory of the instrument itself.

The second defect of the instrument is that its output, the observation, does not in general even represent perfectly the distorted image referred to above. In fact, the output is in general not related in a completely causal way to the input. That is, there are perturbations in the observations which are unpredictable in detail, vary from one observation to the next, and can be understood, if at all, only in terms of the statistical laws which govern them. It is now fashionable to call these perturbations « noise ». To ameliorate the effect of noise is the task of the statistician.

In the diagram above, a duplicated experiment is indicated. This is one in which the phenomenon of interest is absent by design. The resulting observations then calibrate the instrument and, in a statistical sense at least, « calibrate » the noise. Statistical inference or decision results by comparison of the two kinds of observation.

The upper line of the diagram above also represents a mathematical model for the second kind of statistical problem, that of estimating parameters (B, above). In this kind of problem, the phenomenon of interest is the value, θ, of some numerical parameters. It is assumed, or known from the theory of the instrument, that the observation Y is governed by a known statistical law depending in a known way upon the, still unknown, value of θ, thus

$$\text{Probability } \{Y \leqslant y\} = F(y;\theta)\,, \tag{1}$$

where for each θ, $F(y;\theta)$ is a distribution function of y.

From observations of Y, one wishes to estimate a value of θ.

In structure, this problem is not so different from that of testing hypotheses as may at first seem. In the first place, the second line of the diagram above is, in a sense, still present. The knowledge obtained from these « calibrating » experiments is already incorporated in (1), *i.e.*, somehow or other, and ultimately presumably by observation, one has discovered the dependency indicated in (1) between observation Y and parameter θ. In the second place, determining a suitable value for θ is a limiting case, as $N \to \infty$, of testing simultaneously the succession of « neighbouring » hypotheses:

$$\text{hypothesis } k \qquad \frac{k}{N} \leqslant \theta < \frac{k+1}{N}\,,$$

where *e.g.* $-N^2 \leqslant k \leqslant N^2$.

The real difference between problems A and B, testing and estimation, lies in the measures of success used by the statistician. Somewhat loosely, one may say that in testing hypotheses the statistician is interested in how often he accepts the correct hypothesis. All failures to accept the correct hypothesis are equally distressing to him. A characteristic of problems of estimation is that, in general, the parameter being estimated has a quantitative meaning, large errors are considered to be more serious than small ones, and one attempts to estimate, not exactly (since this may be impossible), but in such a way as to minimize some numerical measure of error or of average error.

A practical characteristic of most methods of statistical inference is that they are designed for situations in which there are relatively few basic data, in particular, for problems in which there are only a finite number of random variables. A further characteristic of great theoretical importance is that they are also designed for situations in which the statistician cannot, *a priori*, assume any statistical laws for the underlying phenomenon. For an example, in the agricultural experiment described above, it is unlikely that he would be justified in assuming that « the probability that this fertilizer will have no effect is .3 ». Similarly, in many problems of estimation it is not valid to assume that the parameter θ is itself a random variable.

Some preliminaries, and a description of Shannon's theory.

A proper formulation of the theory of information requires the mathematical apparatus of probability theory. Even a careful descriptive account requires enough of the terminology, to justify its introduction here. Probability begins with an exhaustive listing of all of the « elementary events » which are assumed to be possible. Let W denote the totality of these elementary events. In a simple example, W consists of six elementary events: these are the six possible outcomes in the throw of a single die. It is convenient to call W a « space », and its elements w the points of that space. An *event* is either a point w, or a collection (set) of such points. Thus, in the throw of a die, the occurrence of 1 on the upper face is a « point » of the relevant W; it is also an event. Other events, are, for example, i) W itself (the occurrence of something), or ii) the occurrence of an even number. It is convenient to call the vacuous subset of W also an event.

In all but the simplest problems, it becomes necessary to restrict the events, the subsets of W, for which one calculates probabilities. Theory, however, is useless if there are too few events for which probabilities can be stated. The proper compromise is to consider a family F of events which has three properties of completeness:

i) W itself is in F;

ii) if E is in F, then $W-E$ is in F ($W-E$ is called the complement of E; it consists of all points of W which are not in E);

iii) if E_1, E_2, E_3, ..., are in F, then the union $E_1 v E_2 v E_3 v \ldots$ is in F. (the union $E_1 v E_2 v E_3 v \ldots$ consists of all points of W which are in *some* E_i, $i=1, 2, 3, \ldots$).

A collection F of subsets of W having these three properties is called a Borel field. A Borel field F can be described somewhat loosely, thus: if F contains E_1, E_2, ..., then it also contains any event E which can be described in terms of E_1, E_2, ..., by means of a logically well formulated sentence.

The final element of structure is probability itself. This is a function $P\{E\}$ defined for all events E in F with the three properties:

i) $P\{W\}=1$;

ii) for E in F, $0 \leqslant P\{E\} \leqslant 1$;

iii) if E_1, E_2, ..., are in F and if, when $i \neq j$, E_i and E_j contain no points in common, then

$$P\{E_1 v E_2 v \ldots\} = P\{E_1\} + P\{E_2\} + \ldots$$

(in other words, $P\{E\}$ is additive over mutually exclusive events).

1160

A random variable x is a numerically valued function $x(w)$ defined on W More precisely, it is a function for which such descriptions as « the probability that $x \leqslant a$ » are meaningful. That is, for each number g the set of w such that $x(w) \leqslant g$ must be in F. This is a necessary restriction, but not a serious one. It is convenient also to allow $x(w)$ to be undefined at some points w provided the probability of the set of such w is 0.

For events E and B in F, the conditional probability of E, given B, denoted by $P\{E|B\}$ is by definition

$$P\{E|B\} = \frac{P\{E \wedge B\}}{P\{B\}},$$

when $P\{B\} \neq 0$. It is not definable, in this form when $P\{B\} = 0$. In this definition $E \wedge B$ denotes the intersection of E and B, the set of all w which are both in E and in B. Therefore $P\{E|B\}$ is the relative fraction of the probability in B which is covered by points which are also in E.

Let y be a random variable which assumes only finitely many distinct values. That is, there are numbers $b_1, b_2, \ldots, b_n$ such that the sets B_i, described by

$$B_i = (\text{all } w \text{ such that } y(w) = b_i)$$

exhaust W. Then given an event E, for each B_i such that $P\{B_i\} \neq 0$ (and there must be at last one such, since $1 = \sum P\{B_i\}$) the conditional probability $P\{E|B_i\}$ is defined. This quantity is now a function of i, *i.e.*, a function of y. We use the symbol $P\{E|y\}$ to denote this quantity. For each E it is a random variable, namely, a numerical function of w which takes, for each w in the set B_i, the value $P\{E|B_i\}$. Since the B_i exhaust W, $P\{E|y\}$ is undefined at most over a set of probability 0 (this set being the union of all B_i such that $P\{B_i\} = 0$). Notice that $P\{E|y\}$ is not a general random variable; it is a function of y.

In a similar way, one can define $P\{E|y_1, y_2, \ldots, y_n\}$ given several discrete random variables. From this definition one can then pass by limiting operations to a general $P\{E|y_1, y_2, \ldots\}$, where the y's may be infinite in number, and indeed need not be discrete. In all cases, for fixed E, $P\{E|y_1, \ldots\}$ is a random variable which is, apart perhaps from a set of events having probability zero, a function of the indicated conditioning variables.

Turning now specifically to the probability theory needed for a description of the theory of information, let A be an alphabet *i.e.*, a finite list of symbols $A_1, A_2, \ldots, A_n$. Let W be the collection of all infinite sequences $w_i \ldots, \alpha_{-1}, \alpha_0, \alpha_1, \alpha_2, \ldots$, where each term α_i in the sequence is a letter drawn from A. Consider a set C of sequences described in the following way: C consists of all sequences w such that $\alpha_{t_1} = l_1$, $\alpha_{t_2} = l_2, \ldots, \alpha_{t_n} = l_n$, where the $l_1, \ldots, l_n$ are specified letters, not necessarily distinct, of A. For convenience, call a set C

described in this way a cylinder set. Let F be the smallest Borel field of subsets of W which contains all cylinder sets. Once an alphabet A is given the probability space W is defined and it is a theorem that the Borel field F is then unique. The dependence of W and F upon A will sometimes be indicated by a subscript.

It is also a theorem that to specify a probability for sets of F it suffices to define $P\{C\}$ for cylinder sets C. More exactly if $P\{C\}$ is a true function of cylinder sets C in that $P\{C\}$ is independent of the mode of describing C, and if $0 \leqslant P \leqslant 1$, then the domain of definition of P can be extended by limiting operations to all sets of F, and P defines a probability theorem.

In saying that $P\{C\}$ shall be independent of the mode of describing C, we mean the two things illustrated by the following examples:

$$P\{\alpha_1 = l_1 \text{ and } \alpha_2 = l_2\} = P\{\alpha_2 = l_2 \text{ and } \alpha_1 = l_1\}\,,$$

$$P\{\alpha_1 = l_1 \text{ and } \alpha_2 = l_2\} = \sum_{i=1}^{N} P\{\alpha_1 = l_1 \text{ and } \alpha_2 = l_2 \text{ and } \alpha_s = A_i\}\,.$$

The structure just described: an alphabet A and a probability defined on the resulting F_A, constitutes an information source.

A source is said to be stationary if, for each cylinder set and each $k = 0, \pm 1, \pm 2, \ldots$

$$P\{\alpha_{t_1} = l_1, \ldots, \alpha_{t_n} = l_n\} = P\{\alpha_{k+t_1} = l_1, \alpha_{k+t_2} = l_2, \ldots, \alpha_{k+t_n} = l_n\}\,.$$

That is, the joint distribution of any $\alpha_{t_1}, \ldots, \alpha_{t_n}$ is the same as that of the corresponding $\alpha_{k+t_1}, \ldots, \alpha_{k+t_n}$ and is independent of absolute time.

Consider an alphabet B, in addition to A. An encoder is a sequence Φ_n, $n = 0, \pm 1, \pm 2, \ldots$ of functions from W_A into B with the three properties:

i) for each w, $\Phi_n(w)$ is a letter of B;

ii) for each n, $\Phi_n(w) = \Phi_n(\alpha_n, \alpha_{n-1}, \alpha_{n-2}, \ldots)$ (that is, Φ_n does not depend upon the letters of w which occur after $t = n$);

iii) given $\Phi_n(w)$ for all n, w is uniquely determined.

Most encoders considered in information theory are not stationary, but are periodic in the sense that for some fixed N

$$\Phi_k(l_1, l_2, \ldots) = \Phi_{k+N}(l_1, l_2, \ldots)$$

for all k and for all sequences $l_1, l_2, \ldots,$ of letters of A.

A channel may be thought of as a generalization of an encoder. Different versions of Shannon's fundamental restricted theorem require different definitions of channel, but the general idea can be illustrated thus: let A and B be alphabets and for convenience now denote W_A by X, W_B by Y, F_A by F_X, F_B by F_Y. Consider also the alphabet AB whose letters are the ordered pairs (A_i, B_j) where A_i is a letter of A, B_j a letter of B. W_{AB} consists of all infinite sequences..., $(\alpha_{-1}, \beta_{-1})$, (α_0, β_0), (α_1, β_1), ..., where $x = ...\alpha_{-1}, \alpha_0, \alpha_1, ...$, $y = ... \beta_{-1}, \beta_0, \beta_1, ...$ are respectively drawn from X and Y. Again, we write XY for W_{AB}, and F_{XY} for the corresponding Borel field. In the most general sense, a channel is a function of probability measures on F_X with values which are probability measures on F_{XY}.

More precisely, all channels considered in the literature are of this form: given P, defined on F_X, there is a Q_p, defined on F_{xy}, with three basic properties:

i) if P is stationary, Q_p is also,

ii) if E is an event in F_X, then $Q_p\{E\} = P\{E\}$;

iii) if C is a cylinder set in F_Y which depends only upon β_n for $n \leqslant t$ then

$$Q_p\{C \mid D_1\} = Q_p\{C \mid D_2\} ,$$

for any two cylinder sets D_1, D_2 in F which specify the same values for those α_n for which $n \leqslant t$.

Condition ii) may be illustrated by the following example:

$$Q_p\{\beta_t = \beta_1 \mid \alpha_{t-1} = A_1 \text{ and } \alpha_{t+1} = A_3\} = Q_p\{\beta_t = \beta_1 \mid \alpha_{t-1} = A_1\} .$$

In other words, a channel is a consistent means for inducing a joint distribution Q_p on input and output, given any initial distribution P on the input alone, and, the distribution of the output cannot anticipate the input.

Further elements of structure internal to the channel must be specified before Shannon's fundamental theorem can be proved, but the definition just given is adequate for a discussion of that theorem.

The fundamental quantity of Shannon's theory is a number, called information rate, associated with each stationary information source. For completeness we state one form of its definition, but it suffices for the present discussion merely to know that this number exists. Given P defined on F_X, by definition the information rate of the resulting source is

$$H(x) = \text{average}\left[\sum_{j=1}^{N} - \log P\{\alpha_0 = A_i \mid \alpha_{-1}, \alpha_{-2}, ...\}\right] .$$

Now given a source and a channel, there are three information sources for which rates can be calculated, the input the output and the joint source whose sequences are those of XY. The corresponding rates are $H(X)$ above

$$H(Y) = \text{Average}\left[\sum_{j=1}^{M} -\log Q_p\{\beta_0 = \beta_j \mid \beta_{-1}, \beta_{-2}, \ldots\}\right],$$

and

$$H(X, Y) = \text{Average}\left[\sum_{i=1}^{N}\sum_{j=1}^{M} -\log Q_p\{\alpha_0 = A_i \text{ and } \beta_0 = \beta_j \mid \alpha_{-1}, \beta_{-1},;\ \alpha_{-2}, \beta_{-2};\ \ldots\}\right],$$

where all averages are taken with respect to the probabilities Q_p.

It is a theorem that the quantity $R(X, Y) = H(X) + H(Y) - H(X, Y)$, called the rate from X to Y, is always $\geqslant 0$. This quantity is a function both of the channel and of P. Its maximum, as P is varied, over all stationary measures on F_X, is called the capacity of the channel. For certain restricted kinds of channels then Shannon's fundamental theorem asserts the following:

Given a channel of capacity C and a source of rate $H = H(X)$, if $H < C$, there exists, in a certain limiting sense, an encoder and decoder such that if the text from the source in encoded before presentation to the channel, the decoded output of the channel recreates the text without error.

The « limiting sense » in which this theorem holds is this: in the limit of perfect reception the delay between output and input is infinite, and the encoder is one with an infinite number of distinct functions Φ_n. This limiting case is approached as one insists on smaller and smaller probabilities of error in reception.

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3º Trimestre

The Statistical Theory of Information.

R. M. FANO
Massachusetts Institute of Technology - Cambridge, Mass.

1. – Introduction.

I propose to discuss the foundations of the statistical theory of information to which Dr. McMILLAN referred as the theory of telegraphic communication. This theory, which was formulated by SHANNON in 1948 [1], provides a mathematical framework for the quantitative study of communication processes. The three main topics that I shall cover are best described in terms of the model of communication system illustrated in Fig. 1.

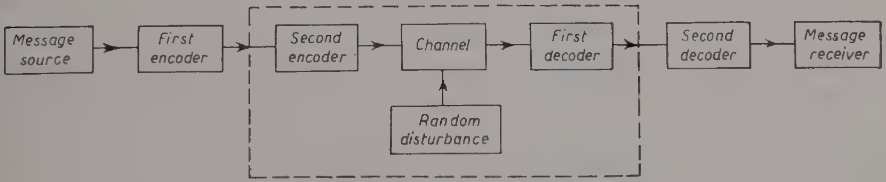


Fig. 1. – Schematic model of communication system.

The message output from the source may be a printed text, a picture, a time function representing the acoustic wave produced by a speaker, the output from a digital computer, etc. The purpose of the first encoder is to represent any such message in a standard form, such as a sequence of binary digits. It seems reasonable to require such a representation to be economical in the sense of using, on the average, as few binary digits as possible.

The purpose of the second encoder is intimately related to the characteristics of the channel, and, in particular, to the fact that the channel is always subjected to random disturbances. We shall assume, for the purposes of our discussion that the channel is discrete in the sense that it accepts as an input only symbols belonging to a specified finite set, such as the letters of the Latin alphabet, and makes available, as an output, symbols belonging also to a specified finite set, not necessarily identical to the first one. Because of random disturbances in the channel the output symbol is not uniquely specified by the input symbol; rather only the probabilities of the different output symbols are functions of the input symbols. Thus, in general, the input symbols cannot be identified with certainty on the basis of the output symbols, and errors will inevitably be committed when any such identification is attempted. The function of the second encoder is to transform the standard representation of the input message into a sequence of symbols which reduces as much as possible the chance of making an erroneous identification of the message after it has been transmitted through the randomly disturbed channel. We shall see that, under certain conditions, it is possible to make the probability of erroneous message identification as small as desired. The function of the first decoder is to identify the sequence of symbols input to the second encoder on the basis of the corresponding sequence of symbols output from the channel. Thus, ideally, the output from the first decoder should be identical to the input to the second encoder. Finally, the function of the second decoder is to reconstruct the original message, on the assumption that its standard representation was correctly identified by the first decoder.

The objective of the overall communication system is to make available to the ultimate receiver a correct replica of the input message, and to accomplish this efficiently, in the sense of using on the average as few channel symbols as possible. It is clear that, in order to be able to speak about the efficiency of the system we must be able to characterize quantitatively what is transmitted through the system, as well as the capability of the specified channel to transmit it. That is, we must define a suitable measure of the extent to which a symbol output from a « black box » specifies the corresponding input simbol, whether the black box be the specified channel or a coding device to be designed. The first topic in this series of lectures is the definition of such a measure and the study of its elementary properties; we shall refer to the object of this « measure » as the *information provided by one symbol about another symbol.*

The measure that we shall define has a great deal of intuitive appeal. We must be very careful, however, to avoid regarding this measure as a fundamental one, just because its properties check so well with our intuitive notion of « information ». Its importance, which is indeed great, stems from the two fundamental theorems stated by C. E. Shannon in 1948 and further refined by Shannon himself and others since that time. The first theorem concerns the operation performed by the first encoder. It states, roughly speaking, that the number of binary digits required, on the average, to represent a message is equal to the average amount of information that must be provided about a message belonging to a specified ensemble in order to identify it uniquely. This theorem constitutes the second main topic in this series of lectures.

The second fundamental theorem concerns the operations performed by the second encoder and by the first decoder. It states that under certain conditions it is possible to encode and decode messages in such a way that the probability of erroneously identifying the message transmitted becomes arbitrarily small. This can be accomplished as long as the average amount of information provided about the message by each symbol input to the channel is smaller than the information capacity of the channel; that is, smaller than the average amount of information that can be provided by each output symbol about the corresponding input symbol. This theorem constitutes the third main topic to be discussed in these lectures.

Because of time limitations we shall confine our discussion to discrete channels, although actual physical channels are more closely represented by continuous models. It will suffice to say here that all the important properties of discrete channels can be generalized to continuous channels.

2. – Definition and elementary properties of a measure of information.

Let $x_1, x_2, \dots x_k, \dots x_{m_x}$ be the points of a discrete space X and $y_1, y_2, \dots y_i, \dots y_{m_y}$ be the points of a discrete space Y. We shall denote with x the variable representing a point of the first space and with y the variable representing a point of the second space. A probability $P(x, y)$ is defined over the product space XY, consisting of the points representing all possible pairs x, y. We may think, for instance, of x as the symbol input to the channel of Fig. 1, and of y as the corresponding output symbol. We may also think of x as the message generated by the source, and of y as one of the symbols output from the first encoder. Our objective is to define a suitable measure of the information provided by the event $y = y_i$, about the event $x = x_k$.

Let us consider this question from the point of view of how well the event $y = y_i$ can identify the event $x = x_k$ in the eyes of a person who can observe only the first event. The « a priori » probability of the event $x = x_k$ is:

$$P(x_k) = \sum_Y P(x_k, y) , \tag{1}$$

where the summation extends over all the values of y. The « a posteriori » probability that the event $x = x_k$ has occured, that is its probability conditioned by $y = y_i$ is:

$$P(x_k/y_i) = \frac{P(x_k, y_i)}{P(y_i)} = \frac{P(x_k, y_i)}{\sum_X P(x, y_i)} \tag{2}$$

where the summation extends over all the values of x. Clearly the effect of the observation of the event $y = y_i$ is to change the probability of the event $x = x_k$ from $P(x_k)$ to $P(x_k/y_i)$. It seems reasonable, therefore, to require that the information provided by the occurence of the event $y = y_i$ about the occurrence of the event $x = x_k$ (for short, provided by y_i about x_k) be a function of these two probabilities. Let us denote this measure of information by $I(x_k; y_i)$. The first requirement on the measure of information to be defined is:

a) $I(x_k; y_i) = F(\varphi, \nu)$, where $F(\varphi, \nu)$ is a once differentiable function of φ and ν and $\varphi = P(x_k)$, $\nu = P(x_k/y_i)$.

We observe, on the other hand, that if the two events involved depend on the occurrence of a third event which has already been observed, consistency requires the a priori and a posteriori probabilities involved in the measure of information to be conditioned by this third event. Let then $z_1, z_2, \dots z_i, \dots z_{m_z}$, be the points of a third space Z represented by a discrete variable z, and assume that a probability $P(x, y, z)$ is defined over the product space XYZ of all triplets x, y, z. Denoting by $I(x_k; y_i/z_j)$ the information provided by y_i about x_k when z_j is given, we require:

b) $I(x_k; y_i/z_j) = F(\varphi, \nu)$ where $\varphi = P(x_k/z_j)$, $\nu = P(x_k/y_i, z_j)$.

Let us consider next the information $I(x_k; y_i, z_j)$ provided by the pair of events $y = y_i$ and $z = z_j$ about the event $x = x_k$. It seems reasonable to require that the measure of this information be the same whether we regard the two events $y = y_i$ and $z = z_j$ as observed simultaneously as a pair or individually in succession. Thus we require that

c) (3) $$I(x_k; y_i, z_j) = I(x_k; y_i) + I(x_k; z_j/y_i) ,$$

Let us consider finally two additional discrete variables ξ and η, statistically independent of x, y

$$P(x, y, \xi, \eta) = P(x, y) P(\xi, \eta) . \tag{4}$$

Because of this statistical independence, the information provided by the events $y = y_i$, $\eta = \eta_h$ about the events $x = x_k$, $\xi = \xi_g$, should be independent of whether we regard the events of each pair as separate events or as forming a single composite event. Thus we must require:

$$d)\ (5) \qquad I(x_k, \xi_g; y_i, \eta_h) = I(x_k; y_i) + I(\xi_g; \eta_h)\,.$$

It can be shown that the four conditions stated above are sufficient to specify uniquely a measure of information. We obtain for this measure

$$(6) \qquad F(\varphi, \nu) = \log \frac{\nu}{\varphi}\,,$$

Thus, for instance, the amount of information provided by y_i about x_k is given by

$$(7) \qquad I(x_k; y_i) = \log \frac{P(x_k/y_i)}{P(x_k)}\,.$$

The derivation of this result is similar to that given by P. M. WOODWARD [2]. The base of the logarithm is arbitrary, and affects only the size of the unit of information. For reasons that will become evident later, the base-two logarithms are most often used, and the name « bit » is employed to denote the unit of information. We shall follow this convention, and all logarithms will be understood to be base two unless otherwise stated.

This measure of information has a number of interesting properties. In the first place, multiplying the numerator and the denominator in Eq. (7) by $P(y_i)$ yields

$$(8) \qquad I(x_k; y_i) = \log \frac{P(x_k, y_i)}{P(x_k)P(y_i)} = \log \frac{P(y_i/x_k)}{P(y_i)} = I(y_i; x_k)\,.$$

Thus the information provided by y_i about x_k is equal to the information provided by x_k about y_i. In other words, the measure is symmetrical in the two events. For this reason it is convenient to refer to it as the « *mutual information* » between the two events. Clearly, it is a measure of the extent to which the two events are more likely to occur together than if they were statistically independent.

The mutual information $I(x_k; y_i)$ becomes a maximum for a fixed $P(x_k)$ when $P(x_k/y_i) = 1$, that is, when x_k is uniquely specified by y_i. This maximum value is equal to

$$(9) \qquad I(x_k) = -\log P(x_k)\,.$$

We shall refer to $I(x_k)$ as the « *self information* » of the event $x = x_k$. It represents the amount of information that must be provided about this event by some other event in order that the former be uniquely specified. Conversely, if we regard the event $x = x_k$ as providing information about some other event, its self information represents the maximum amount of information that it can provide about the other event. In general

$$(10) \qquad I(x_k; y_i) \begin{cases} \leqslant I(x_k)\,, \\ \leqslant I(y_i)\,. \end{cases}$$

It is important to observe that $I(x_k; y_i)$ may be negative as well as positive. It is negative when the probability of occurrence of the pair of events $x = x_k$, $y = y_i$ is smaller than if the variables x and y were statistically independent. On the other hand the average value (expectation) of $I(x_i; y_i)$ over the set of events X can be shown to be non-negative, that is,

$$(11) \qquad I(x; y_i) = \sum_X P(x/y_i) I(x; y_i) \geqslant 0\,.$$

This result checks with our intuitive notion that, if x and y are statistically related, the event $y = y_i$ must provide on the average a positive amount of information about x.

The average value of the self-information associated with x is given by

$$(12) \qquad H(X) = \sum_X P(x) I(x) = -\sum_X P(x) \log P(x) \geqslant 0\,.$$

Because of the form of this expression, it is customary to refer to it as the *entropy* of the ensemble of events X. It can be interpreted as the amount of information that must be provided, on the average, to identify one of the events of the ensemble. It can also be interpreted as the maximum amount of information that can be provided on the average by one of the events of the ensemble X about some other event.

The entropy $H(X)$ is a function of the probability distribution $P(x)$. It can be readily shown that it becomes a maximum when

$$(13) \qquad P(x) = \frac{1}{m_x}\,,$$

for all the m_x events of the set, that is when the events are equiprobable. The maximum value is:

$$(14) \qquad H(X)_{\max} = \log m_x\,.$$

The conditional entropy

$$H(Y/X) = \sum_X \sum_Y P(x, y) I(y/x) = -\sum_X \sum_Y P(x, y) \log P(y/x) , \tag{15}$$

represents the average amount of information that must be provided in order to specify the event y when the event x is known. The average amount of information that must be provided in order to specify both events is given by the joint entropy:

$$H(X, Y) = \sum_X \sum_Y P(x, y) I(x, y) = -\sum_X \sum_Y P(x, y) \log P(x, y) . \tag{16}$$

We obtain with the help of Eq. 12 and 15

$$H(X, Y) = H(X) + H(Y/X) . \tag{17}$$

It can be shown, in addition, that

$$H(Y/X) \leqslant H(Y) , \tag{18}$$

in words, the amount of information that must be provided, on the average, in order to specify y can only be decreased by the knowledge of x.

3. – The first fundamental theorem.

We saw that the function of the first encoder in Fig. 1 is to provide a representation of the input message in the form of a sequence of symbols selected from a specified alphabet. The character of this encoding operation depends, among other things, on the form in which the message is generated by the source. The essence of the problem, however, can be stated in simple terms as follows.

Let us consider an ensemble of M messages, $u_1, u_2, ..., u_k, u_M$, with corresponding transmission probabilities $P(u_1), ..., P(u_M)$. Let us suppose, also, that each message must be represented, for transmission purposes, by means of a sequence of symbols (code words) selected from an alphabet with D symbols, where $D < M$. We wish to inquire about the minimum number of symbols required, on the average, to specify one of the messages of the ensemble.

It was pointed out in the preceding section that the maximum amount of information that can be provided on the average by one symbol is equal to log D, the capacity of the alphabet, and that this maximum amount is actually provided when the symbols of the alphabet occur with equal propababilities. This suggests that we should construct the code words in such a way that at each position in them the different symbols of the alphabet will occur equiprobably, and independently of the preceding symbols. The significance of this statement can be best understood in terms of the following examples of binary code words.

The 8 messages in Fig. 2 have the same transmission probability $P(u) = 2^{-3}$. The first binary digit [*] is 0 for the first four code words, and 1 for the remaining four code words. Thus the probability that the first digit be 0 is exactly equal to the probability that the first digit be 1; in other words, the two sets of messages separated by the first digit are equally probable. The second digit divides again the message space into two equiprobable sets, with 0 being assigned to messages u_1, u_2, u_5, u_6, and 1 being assigned to messages u_3, u_4, u_7, u_8, furthermore, these two sets of messages intersect the two sets formed by the first digit in such a way as to yield four equiprobable subsets, namely $u_1, u_2; u_3, u_4; u_5, u_6; u_7, u_8$. Thus the probabilities that the second digit be a 0 and that it be a 1 are not only equal, but also independent of the first digit. The third digit divides once more the message space into 2 equiprobable sets, which in turn divide each of the above four subsets into two equiprobable parts, each consisting of a single message. Thus the third digit too is independent of the preceding digits as well as equiprobable. Since the three successive digits are used at full capacity, they must provide together an amount

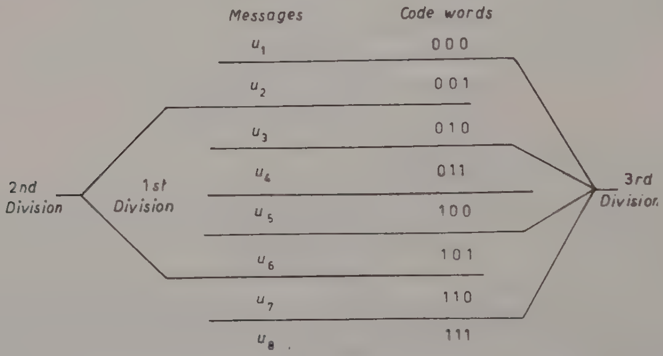


Fig. 2. – Optimum set of code words for equiprobable messages.

[*] Because of the frequent use of binary digits, the use of the term « digit » is restricted hereafter to the binary case. The term « symbol » is used in the general case of a D-alphabet.

of information equal to 3 binary units (bits); this is just equal to the entropy of the message ensemble, that is to the amount of information that must be provided on the average to identify a message of the ensemble.

Messages	Probabilities	Code words
u_1	0.25	0 0
u_2	0.25	0 1
u_3	0.125	1 0 0
u_4	0.125	1 0 1
u_5	0.0625	1 1 0 0
u_6	0.0625	1 1 0 1
u_7	0.0625	1 1 1 0
u_8	0.0625	1 1 1 1

Fig. 3. – Optimum set of code words.

In the case of Fig. 3, the messages are no longer equiprobable, but their probabilities are still of the form

$$P(u_k) = 2^{-n_k}\,, \tag{19}$$

where n_k is an integer. Again the first binary digit divides the message space into two equiprobable sets, which, however, do not contain the same number of messages. The second digit divides each of these two sets into two equiprobable subsets. Two of the resulting subsets contain a single message, while the other subsets are further divided by the third and fourth digit into equiprobable parts until all messages are singled out.

It is clear by inspection that each digit is independent of all preceding digits and that it may be a 0 or a 1 with equal probabilities. Furthermore all digits of a code word are uniquely specified by the corresponding message, so that the mutual information between each digit and the message is equal to the self information of the digit. Thus each digit is used at full capacity, that is, it provides as much information as possible about whatever message is being transmitted. As a matter of fact, since each digit, whether a 0 or a 1, occours with probability $\frac{1}{2}$, it provides in all cases exactly one unit of information about the corresponding message. It follows that the sum of the self informations of the individual digits of each code word must be equal to the number of digits in the code word. On the other hand this sum must also be equal to the self information of the corresponding message, because each code word is uniquely specified by the corresponding message. We may conclude, therefore, that the number of digits in each code word must be equal to the self information of the corresponding message, that is to the integer n_k in Eq. (19). This is actually the case in Fig. 3, as it can be readily checked by inspection. Furthermore, the integers n_k satisfy the equation

$$\sum_{k=1}^{M} 2^{-n_k} = \sum_{k=1}^{M} P(u_k) = 1 \, . \tag{20}$$

The average number of symbols per code word can be readily computed. We have

$$N = \sum_{k=1}^{M} P(u_k) n_k = 2.75 = -\sum_{k=1}^{M} P(u_k) \log P(u_k) \, , \tag{21}$$

which is just the entropy of the message ensemble. This result is in agreement with the intuitive notion that, since each digit is used at full capacity (it contributes one unit of information) the average number of digits per code word must be equal to the amount of information required, on the average, to identify a message (the entropy of the message ensemble). The same reasoning suggests also that the set of code words illustrated in Fig. 3 is optimum, in the sense that no set of code words could yield a smaller value for the average number of digits per code word. This is actually the case, as shown below.

The idea of constructing code words by successive divisions of the message space into equiprobable sets and subsets applies to the case of an arbitrary alphabet as well as to that of a binary alphabet; one must only divide into D equally probable subsets instead of just two. The technique fails, however, as one would expect, when the message probabilities are not negative powers of D (the number of symbols in the alphabet) because it becomes impossible at some point to make equiprobable divisions. Still, one may try to make the divisions as equiprobable as possible, thereby hoping to keep the average number of symbols per message as small as possible.

Upper and lower bounds on the minimum average number of code symbols per message can be obtained with the help of the following theorem due to L. Kraft. Before this theorem can be stated, however, we must discuss an important condition that must be satisfied by the code words assigned to the messages. It is obvious that in order for each message to be *uniquely* specified by the corresponding code word, no two code words can be identical. Furthermore, no code word can be a continuation of a shorter code word; for instance the code words 0 0 1 and 0 0 1 0 cannot be present in the same set. This follows from the fact that two such code words can be distinguished only because empty spaces are present in one where symbols are present in the other. Basing the identification of the code words on such a difference would be equivalent to using an empty space as an additional symbol, thereby increasing the size of the coding alphabet. We can now state the theorem mentioned above.

Let $n_1, n_2, \dots n_k, \dots n_M$ be a prescribed set of M positive integers. The inequality

$$\sum_{k=1}^{M} D^{-n_k} \leqslant 1 \tag{22}$$

is a necessary and sufficient condition for the existence of a set of M different code words employing an alphabet with D symbols, whose lengths are equal to the prescribed integers, and none of which is a continuation of a shorter one. The proof of this theorem can be found in reference [3].

The average number of symbols per message is, by definition,

$$N = \sum_{k=1}^{M} P(u_k) n_k \,. \tag{23}$$

We wish to find upper and lower bounds to the minimum value of N for any given message ensemble. A lower bound can be readily found by relaxing the condition that the lengths of the code words be integers. Under these conditions we can minimize the right-hand side of Eq. (23) with respect to the variables $n_1, n_2, \dots n_k, \dots n_M$, subject to the constraint imposed by Eq. (22). The resulting minimum value is clearly a lower bound for N. We obtain for the optimum value of n_k

$$n_k = \frac{-\log P(u_k)}{\log D} = \frac{I(u_k)}{\log D}\,, \tag{24}$$

and for the desired lower bound

$$N_{\min} = \sum_{k=1}^{M} P(u_k) \frac{I(u_k)}{\log D} = \frac{H(U)}{\log D}\,. \tag{25}$$

It is interesting to note that this bound is equal to the entropy of the message ensemble divided by the capacity of the alphabet, as suggested by the above examples. Furthermore, the optimum length of the code word corresponding to each message is just equal to the self information of the message divided by the capacity of the alphabet. Thus if the self information of each message happens to be an integral multiple of the alphabet capacity, the optimum word lengths are integers, and a corresponding optimum set of code words is insured by the theorem stated above. For instance, in Figs. 2 and 3 all the message self-informations were integers, and, as a result, the average number of binary digits per message could be made equal to the entropy of the message ensemble. Thus our intuition was correct in suggesting that the sets of code words of Figs. 2 and 3 were optimum.

An upper bound to the required average number of symbols per message can be obtained as follows. We observe, first of all, that the only reason why it is not possible in general to make the average number of symbols per message equal to the optimum value given by Eq. (18), is that the ratios $I(u)/\log D$ are not usually equal to integers. It would seem reasonable, in such cases, to make each n_k equal to the integer n_k^* just larger than the corresponding ratio, so that

$$\frac{I(u_k)}{\log D} \leqslant n_x^* < \frac{I(u_k)}{\log D} + 1 . \tag{26}$$

On the other hand this inequality implies

$$\sum_{k=1}^{M} D^{-n_k^*} \leqslant \sum_{k=1}^{M} 2^{-I(u_k)} = \sum_{k=1}^{M} P(u_k) = 1 . \tag{27}$$

Thus, in view of the theorem proved in the preceding section, it is always possible to construct a set of code words with the lengths specified by the inequality (26). The resulting average number of symbols per message is given by

$$N^* = \sum_{k=1}^{M} n_k^* P(u_k) < 1 + \frac{H(U)}{\log D} . \tag{28}$$

Finally combining this upper bound with the lower bound given by Eq. (25) yields

$$\frac{H(U)}{\log D} \leqslant N \leqslant 1 + \frac{H(U)}{\log D} , \tag{29}$$

where N is the required average number of symbols per message.

The per cent difference between the upper bound and the lower bound becomes negligibly small for large values of $H(U)/\log D$. Let us suppose, for instance, that the messages consist of segments of sequences of independent x-symbols, belonging to ensembles having the same entropy $H(X)$. The entropy of the message ensemble consisting of all possible segments of length n is then equal to

$$H(U) = nH(X) . \tag{30}$$

Then, substituting this equation for $H(U)$ in Eq. (29) and dividing by n yields

$$\frac{(X)H}{\log D} \leqslant \frac{N}{n} \leqslant \frac{1}{n} + \frac{H(X)}{\log D} . \tag{31}$$

It follows that the minimum average number of code symbols per message symbol can be made as close as desired to $H(X)/\log D$ by making the message length n sufficiently large. Essentially the same result is obtained when the successive x-symbols constituting a message are not statistically independent.

The above result constitutes the first fundamental theorem. Its importance lies in the fact that it provides a direct operational meaning for the entropy of a message ensemble as the minimum number of binary digits required on the average to encode a message. We might also say, from a broader point of view, that the theorem provides substantial evidence for the soundness of the postulates on which the statistical theory of information is based.

4. – The second fundamental theorem.

We saw in the preceding section that it is possible to represent the messages of a given ensemble by means of sequences of symbols in such a way that each successive symbol is used at close to full capacity. This implies that the output of the first encoder can be regarded for all practical purposes as a sequence of independent, equiprobable symbols. For the sake of simplicity we shall assume, in what follows, that these symbols belong to a binary alphabet, and refer to them as binary digits. Since they can be regarded as independent and equiprobable, each particular sequence of n such digits will occur with probability 2^{-n}.

The second fundamental theorem in its simplest form concerns the problem of transmitting sequences of independent, equiprobable digits through a discrete channel with specified characteristics. Thus it involves the part of the system of Fig. 1 enclosed by a dotted line, namely the second encoder, the channel, and the first decoder. The goal is, of course, to reproduce correctly at the output of the first decoder the sequence of binary digits input to the second encoder, and employ for this purpose as few channel symbols as possible.

Let us consider first the transmission properties of the channel. If we represent with x the symbol input to the channel and with y the corresponding output symbol, the channel is defined by the set of conditional probabilities $P(y/x)$. We shall limit our discussion to stationary channels without memory that is to channels for which the values of $P(y/x)$ are fixed and independent of the preceding input and output symbols.

Let us denote by $P(x)$ the probability of the symbol x. The average amount of information provided by y about x is obtained by averaging the mutual information between x and y over all pairs x, y. We obtain

$$I(X;Y) = \sum_X \sum_Y P(x,y) I(x;y) = \sum_X \sum_Y P(x,y) \log \frac{P(x,y)}{P(x)P(y)} . \tag{32}$$

This expression can also be written in terms of entropies in the three alternate forms

$$I(X;\, Y) = H(X) + H(Y) - H(X,\, Y)\,, \tag{33}$$

$$I(X;\, Y) = H(X) - H(X/Y)\,, \tag{34}$$

$$I(X;\, Y) = H(Y) - H(Y/X)\,. \tag{35}$$

Furthermore, it can be readily shown that

$$I(X;\, Y) \geqslant 0\,, \tag{36}$$

The right-hand sides of Eqs. (33), (34) and (35) provide three different interpretations for the average value of the mutual information. According to Eq. (33), it can be interpreted as the difference between the average amount of information necessary to specify x and y separately (as if they were independent) and the amount necessary to specify them together as a pair. In other words, $I(X;\, Y)$ is a measure of the statistical constraint between x and y; it vanishes when x and y are statistically independent. On the other hand, Eq. (34) indicates that $I(X;\, Y)$ is the difference between the average amounts of information necessary to specify x before and after the reception of y. The conditional entropy $H(X/Y)$ is often referred to as the « equivocation » because it represents the uncertainty about x that remains after the reception of y. Finally, Eq. (35) expresses $I(X;\, Y)$ as the difference between the average amount of information that y is capable of providing, and the amount necessary to specify the channel disturbance. The conditional entropy $H(Y/X)$ is sometimes referred to as the « noise entropy ».

Our first objective is to determine the information capacity per channel symbol. We observe, in this regard, that in general, each output symbol provides information not only about the corresponding input symbol, but also about all the preceding input symbols. It can be shown, however, that, for given symbol transmission probabilities the average information provided by each received symbol is a maximum when successive transmitted symbols are statistically independent. Thus, in computing the information capacity of the channel successive symbols can be regarded as statistically independent. It should be stressed however, that this is true only for channels without memory, that is only when the values of the conditional probability $P(y/x)$ are independent of all preceding symbols.

The average mutual information given by Eq. (32) is a function of the values of the transmission probabilities $P(x)$ as well as of the characteristics of the channel, represented by the values of the conditional probability $P(y/x)$. Let

$$C = \max_{P(x)} I(X;\, Y)\,, \tag{37}$$

be the maximum average value of the mutual information between x and y, obtained by varying the symbol transmission probabilities. In view of the preceding argument, C is an upper bound to the average amount of information that can be provided by each received symbol about the corresponding transmitted symbol and all preceding symbols. We shall ıefer to C as the channel capacity.

Let us evaluate, as a simple example, the capacity of a binary symmetric channel in which the digits 0 and 1 are received incorrectly with probability p, and correctly with probability q,

$$P(y=0/x=0)=P(y=0/x=0)=q=1-p\,, \tag{38}$$

$$P(y=0/x=1)=P(y=1/x=0)=p\,. \tag{39}$$

We obtain for the noise entropy

$$H(Y/X)=-\sum_X\sum_Y P(x)P(y/x)\log P(y/x)=-p\log p-q\log q\,, \tag{40}$$

Because of the symmetry of the channel this expression is independent of the transmission probabilities. It follows from Eq. (35) that the maximum value of $I(X;Y)$ can be obtained simply by maximizing $H(Y)$. In our particular case, the maximum value of $H(Y)$ is equal to one, and it is achieved when the digits 0 and 1 are received with equal probabilities. This implies, in turn, that the two digits are also transmitted with equal probabilities. Thus, we obtain for the capacity of the binary symmetric channel

$$C=1+p\log p+q\log q\,. \tag{41}$$

Base two logarithms are used in accordance with the convention established in Sec. **2**. Clearly, C is equal to unity for $p=0$, that is when the probability of incorrect reception is equal to zero; it decreases monotonically with increasing p and vanishes for $p=\frac{1}{2}$, that is when the received digit is statistically independent of the transmitted digit.

It should be stressed that while the evaluation of the capacity is very simple in the case of a binary symmetric channel, il may become very involved in other cases. The source of difficulty is that the maximization of $I(X;Y)$ must be carried out under the constraint that the values of the transmission probabilities be non-negative numbers.

Let us turn our attention next to the encoding and decoding operations performed by the second encoder and by the first decoder in Fig. 1. For the sake of simplicity, we shall limit most of our discussion to the case of a binary symmetric channel. In such a case the encoder and the decoder transform sequences of binary digits into other sequences of binary digits. The overall objective is to reproduce correctly at the output of the first decoder the sequence of binary digits input to the second encoder.

We observe, first of all, that the amount of information necessary to specify each of the digits input to the second encoder (message digits) is equal to one unit, in view of the fact that the digits are, by assumption, equiprobable and statistically independent. On the other hand the channel capacity is smaller than unity. It follows that the reception of a channel digit cannot provide enough information to specify uniquely one message digit. More precisely, the correct reproduction at the output of the first decoder of N_i message digits certainly requires the transmission of N channel digits where

$$N > \frac{N_i}{C}. \tag{42}$$

We shall refer to

$$R = \frac{N_i}{N} < C\,, \tag{43}$$

as the rate of transmission per channel digit.

It should be clear that, since there are 2^{N_i} possible sequences of N_i message digits and 2^N possible sequences of N channel digits, there is a great deal of freedom in the assignment of channel sequences to message sequences. The coding problem consists of the selection of a set of 2^{N_i} channel sequences out of the possible 2^N sequences, in such a way as to maximize the probability that the first decoder will reproduce correctly the message sequence. The overall operation of the part of the system enclosed by dotted lines in Fig. 1 is illustrated schematically in Fig. 4.

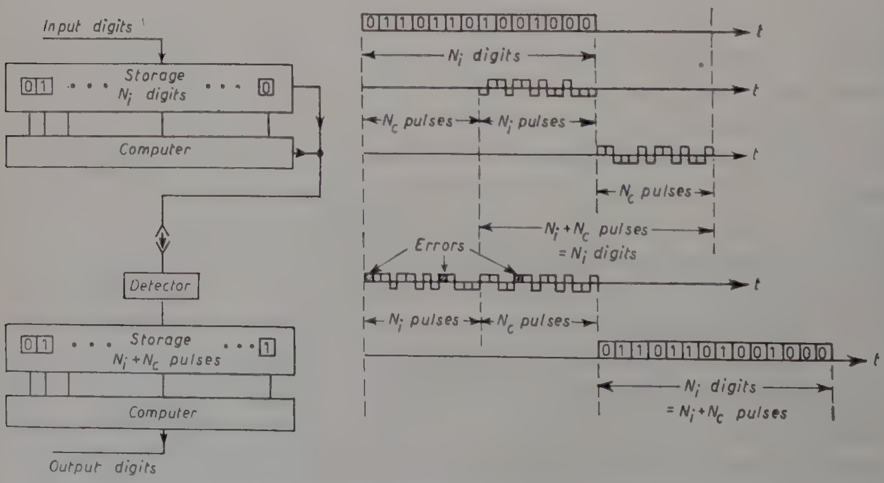


Fig. 4. -- Schematic illustration of coding and decoding operations for trasmission through a binary symmetric channel.

The binary channel is shown in Fig. 4 as a pulse communication system in which a positive pulse corresponds to the digit 1 and a negative pulse to the digit 0. The pulse detector at the receiver is considered as part of the given channel, and it is assumed that a pulse of either polarity has a probability p of being mistaken for a pulse of the opposite polarity. The second encoder consists of the storage device and the computer shown in the upper part of the figure, while the first decoder consists of the storage device and the computer shown in the lower part of the figure.

The encoding operation is performed as follows. The input message digits are stored in blocks of N_i digits. These digits are then transformed into an equal number of corresponding pulses. An additional number N_c of checking pulses are generated by the computer from the N_i message digit, according to suitably selected rules. The entire sequence of

$$N = N_i + N_c\,, \tag{44}$$

pulses is then transmitted through the channel. In Fig. 4 N_c is equal to N_i so that the transmission rate per pulse is equal to one half. This particular method of assigning sequences of pulses to sequences of message digits can be shown to be sufficiently general in the sense that it does not limit in any substantial manner the overall performance of the system.

The sequence of N pulses output from the detector will include, in general, pulses with incorrect polarity. The function of the computer is to select from the set of 2^N sequences of pulses that correspond to message sequences, the one that differs from the received sequence in the least number of pulses. Once this pulse sequence has been determined, the computer generates the corresponding message sequence which constitutes the output of the first detector in Fig. 1. It can be shown that this decoding procedure minimizes the probability that any one of the N_i output message digits be incorrect. We shall denote with P_e this probability of error.

Unfortunately our present knowledge of the optimum rules for computing the additional N_c pulses is very limited. Dr. SLEPIAN will speak (*) in some detail about this problem. On the other hand, P. ELIAS [4] has been able to compute upper and lower bounds to the minimum probability of error that can be achieved in principle. These bounds can be written in the form

$$K_2\, 2^{-NE_2} \leqslant P_e \leqslant K_1\, 2^{-NE_1}\,, \tag{45}$$

where E_1 and E_2 are functions of the channel capacity C and of the rate of transmission R, *but are independent of* N, and K_1 and K_2 are quantities that vary very slowly with N. The quantities E_1 and E_2 are positive for $R < C$

(*) See this issue, page 373.

and vanish for $R = C$. Thus it is possible to make the probability of error as small as desired by increasing the length of the message and channel sequences while keeping their ratio,

$$R = \frac{N_i}{N} < C \,, \tag{46}$$

constant. In other words, it is possible by proper encoding to transmit at any finite rate smaller than the channel capacity with a vanishingly small probability of error per message sequence. This is the second fundamental theorem for the binary symmetric channel.

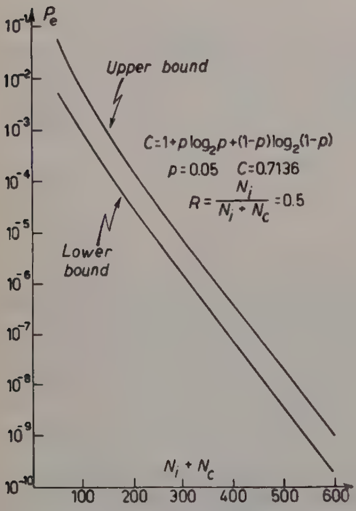


Fig. 5. – Upper and lower bounds to the probability of error for a message consisting of N_i binary digits and encoded into $N_i + N_c$ binary pulses.

The curve shown in Fig. 5 illustrate the behavior with N of the upper and lower bounds for a channel with $p = 0.05$ and for a transmission rate $R = 0.5$. In this particular case the exponents E_1 and E_2 are equal as indicated by the fact that the two curves become parallel straight lines for large values of N.

The functional dependence of E_1 and E_2 on the pulse error probability p that characterizes the channel, and on the transmission rate R is best expressed in terms of the parameter r defined by

$$R = C(r) = 1 + r\log r + (1-r)\log(1-r) \,. \tag{47}$$

This function whose behavior is illustrated in Fig. 6, becomes identical with the channel capacity when we set $r = p$. Let us define, in addition, the quantity

$$T_p(r) = 1 + r\log p + (1-r)\log(1-p) \,. \tag{48}$$

This quantity is a linear function of r for a given p and becomes equal to the channel capacity for $r = p$,

$$T_p(r = p) = C(r = p) = C \,. \tag{49}$$

It is represented in Fig. 6 as a straight line tangent to the curve $C(r)$ at the point $r = p$.

The exponent E_1 is given by

$$E_1 = C(r) - T_p(r) \,, \tag{50}$$

which, in Fig. 6, is the vertical distance between the curve and the tangent for the value of r for which $C(r)$ is equal to the desired rate R. The expression for the exponent E_2 depends on wether, for the desired rate R, the parameter r is larger or smaller than the critical value r_c defined by

$$\left[\frac{r_c}{1-r_c}\right]^2 = \frac{p}{1-p}\,. \tag{51}$$

For r smaller than r_c, E_2 is equal to E_1,

$$E_2 = E_1 = C(r) - T_p(r)\,; \qquad r < r_c\,. \tag{52}$$

On the other hand,

$$E_2 = C(r) - [2C(r) + + T_p(r_c) - 2C(r_c)] < E_1\,; \quad r \geqslant r_c\,. \tag{53}$$

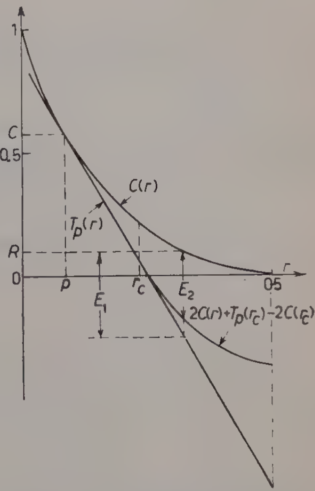


Fig. 6. – Graphical determination of E_1 and E_2.

The curve representing the term in square brackets is tangent to the straight line representing $T_p(r)$, as illustrated in Fig. 6, so that the curve can be regarded as a continuation of the straight line for $r \geqslant r_c$. Thus the exponent E_2 can be interpreted as the vertical distance between the curve $C(r)$ and the new curve.

It is important to note that, since $E_1 = E_2$ for $r < r_c$, the upper and lower bounds have in this region the same exponential behavior with N. As a matter of fact the proportionality factors K_1 and K_2 are both proportional to $N^{-\frac{1}{2}}$, so that their ratio is independent of N. Thus, in this region, the probability of error is effectively bracketed by the two bounds, as illustrated in Fig. 5. On the other hand, for $r \geqslant r_c$, $E_2 < E_1$ and the two bounds diverge exponentially. Thus the probability of error is only roughly bracketed by the bounds for low transmission rates.

The second fundamental theorem is the cardinal result in the statistical theory of information. C. E. SHANNON was the first to show in 1948 [1] that, in any channel without memory, the probability of error can be made as small as desired by proper encoding, for any transmission rate smaller than the channel capacity. This result was refined and extended by A. FEINSTEIN in 1954 [5]. He proved, in particular, that the probability of error can be made to vanish exponentially with increasing message length. The exponential upper bound on the probability of error for the binary symmetric channel was developed independently by P. ELIAS and C. E. SHANNON. More recently. C. E. SHANNON developed [6] a similar upper bound for arbitrary discrete channels without memory.

REFERENCES

[1] C. E. Shannon: *A Mathematical Theory of Communication*, in *Bell Syst. Techn. Journ.* (July and October 1948).

[2] P. M. Woodward and I. L. Davies: *Proc. Inst. Electr. Eng.*, (Part III), **99**, 45 (1952).

[3] A. Feinstein: *Foundation of Information Theory* (New York, 1958).

[4] P. Elias: *Coding for Noisy Channels*, in *IRE Convention Record*, Part 4, 1955.

[5] A. Feinstein: *A New Basic Theorem of Information Theory*, in *IRE Trans. Profess. Group Inform. Theor.* (Sept. 1954).

[6] C. E. Shannon: *Certain Results in Coding Theory for Noisy Channels*, in *Information and Control* (Sept. 1957).

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3° Trimestre

Coding Theory.

D. Slepian

Bell Telephone Laboratories - Murray Hill, N. J.

We have had the pleasure of hearing Drs. Mc Millan and Fano present in some detail the discrete case of the mathematical theory of information originally developed by Claude Shannon. The portion of the theory that they have discussed is now rather fully developed. The obvious and important questions have been asked and well answered. There is of course yet work to be done, extensions and generalizations to be made, new corners to be explored. But by and large, this portion of the theory is in good shape. Drs. McMillan and Fano have been kind enough to leave for me to discuss that part of the theory about which no one knows anything. I can, therefore, safely feel well qualified to speak.

1. – I shall discuss what is known as « the coding problem » of information theory. Actually there are two very distinct coding problems. One concerns information sources and may be called the problem of redundancy removal. The other concerns channels and may be called the problem of combatting noise. Both problems can be further subdivided into two cases as shown in Table I.

Table I. – *The two coding problems.*

	Source	Channel
	Remove redundancy	Combat noise
Discrete	Codes of Shannon, Fano, Huffman, Schutzenberger, etc.	?
Continuous	Work on speech and television signals ?	?

The discrete case of the source coding problem is well in hand now. The problem is the canonical representation in most efficient form of the encoded version of the message given the size of the new alphabet. This is the encoding problem that Prof. FANO has spoken about. The entropy of the source in bits measures the average number of zeros and ones per letter of the original message necessary for this efficient encoding. Coding techniques by SHANNON, HUFFMAN and FANO tell us explicitly how to perform the encoding. Researches by SCHUTZENBERGER, SARDINAS and others have answered many questions about the nature of the coders. Here our knowledge is rather complete.

It has been mentioned several times that much of the theory presented for the discrete case can be extended to continuous messages. When we attempt to make this extension for the redundancy removal problem we encounter a number of complications of a non-trivial sort. Consider for example an ensemble of continuous messages representing speech. It is clear that not all the fine detail of the waveforms is pertinent to speech. We feel that the message in this form has much redundancy, and we naturally seek a minimal encoding of the speech ensemble into sequences of zeros and ones. But minimal in what sense? Here is the difficulty. When we consider the efficient representation of discrete messages by zeros and ones, we require our encodings to be uniquely decipherable into the original messages without errors. When we encode speech into sequences of zeros and ones we require that they be capable of decipherment into speech sounds in some sense as acceptable to the listener as the original uncoded version. Our constraints on the minimization problem, then, are subjective in nature, and until such time as they can be stated in mathematical form much of the work on efficient encoding of speech must be of an experimental nature. A great deal of research is going on in laboratories throughout the world on the redundancy removal encoding problem both for speech and television. Time does nor permit further discussion of it here.

As regard the second coding problem, that of combatting noise on channels, our knowledge to date is very fragmentary indeed. The literature on this subject has grown markedly in the last several years, however, and I shall have little trouble filling the remainder of my talks with selected findings from a few of these papers.

The fundamental theorem states that given a channel with capacity C and an information source producing messages with entropy H, there exists an encoder and a decoder such that messages from the source can be encoded, transmitted over the channel, and decoded with an arbitrarily small probability of error provided only that that $H < C$. If $H > C$, it is impossible to find such an encoder and decoder: the probability of error will be bounded away from zero and we cannot make it arbitrarily small. This theorem is an existence theorem. As proved today, it does not show us explicitly how to construct such encoders and decoders.

From the practical point of view, this theorem contains the golden fruit of the theory. It promises us communication in the presence of noise of a sort that was never dreamed possible before: perfect transmission at a reasonable rate despite random perturbations completely outside our control. It is somewhat disheartening to realize that today, ten years after the first statement of this theorem, its content remains only a promise, that we still do not know in detail how to achieve these results for even the most simple non-trivial channel.

In order to understand better both the theorem, its proof, and our difficulties in achieving the results promised by the theorem, I am going to restrict my attention for most of the remainder of my talks to one particularly simple channel. The problem for more complicated channels differs only in detail but not in overall character.

Let us direct our attention to the binary symmetric channel already discussed by Dr. FANO (see Fig. 1).

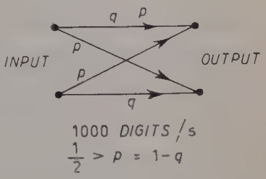


Fig. 1. – Binary symmetric channel.

For ease in talking about the channel, we shall suppose it can handle binary digits at the rate of 1000 per second so that we shall measure its capacity in bits/second rather than in the somewhat more awkward units bits/binary digit transmitted. The capacity then is $C = 1000\,(1 + p \log p + q \log q)$ bits/s. with logarithms taken to base 2. A plot of C versus p is shown in Fig. 2. If, for example, $p = .001$, a not unrealistic value, it is found that $C = 988$, so that according to the fundamental theorem, we should be able to transmit messages with any entropy $H < 988$ bits/second over the channel with as little error as we desire. How shall this be done?

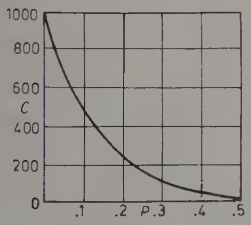


Fig. 2. – Capacity of binary symmetric channel.

For simplicity, let us take as our message source an experimenter tossing a fair coin at some fixed rate, *i.e.*, some fixed number of tosses per second. If we denote heads by one and tails by zero, the experimenter generates a sequence of binary digits whose entropy rate in bits/second is equal to the rate at which the coin is tossed. We wish to transmit the results of this experiment over the channel to a destination.

If we encode the message digits one by one directly into the channel digits, we can transmit up to 1000 digits per second, but there will be probability p that each received digit is in error. One obvious way to decrease this error probability in digits passed on to the destination is to require the experimenter to slow down. If, for example, he tosses his coin at the rate 333 $\frac{1}{3}$ tosses per second, we have time to repeat each digit two additional times in transmission (see Fig. 3).

1 0 1 1 0 1	experimenter's message
1 1 1 0 0 0 1 1 1 1 1 1 0 0 0 1 1 1	transmitted sequence
0 1 1 0 1 0 1 1 1 1 0 1 0 0 1 0 1 1	received sequence
1 0 1 1 0 1	decoded sequence

Fig. 3.

The received digits will in general differ from those transmitted. If we break up the received digits into blocks of three, however, and decode by the majority rule « one if more ones than zeros in a block, zero if more zeros than ones in a block », we will correct all errors in transmission provided only that two or three errors never occur in one block. An elementary calculation shows that the error probability in the digits passed on to the destination is now 10^{-5} instead of 10^{-3} as would result if the experimenter's digits were transmitted without repetition.

The error probability can clearly be reduced further by the same technique. If we require the experimenter to toss the coin at the slow rate of 200 tosses per second, we have time to repeat each digit 5 times in transmission. If we break up the received digits into blocks of length 5 and again use a majority rule, an error probability of 10^{-8} results.

We can of course carry this scheme further and further. The evident disadvantage is that as we decrease the error probability more and more, we require the experimenter to signal more and more slowly. Indeed in the limit of zero error probability we require him to stop generating messages altogether.

Now the main point of Shannon's fundamental theorem is precisely that this slow down is *not* necessary to achieve arbitrarily small error probability. In the case at hand, it asserts that our experimenter can toss the coin at any fixed rate less than 988 tosses per second, and by being sufficiently clever we can deliver the results to a destinations with as little error as we choose.

To gain insight into how this might be done, let us suppose the coin is tossed at the rate 500 tosses/second. We then have time to transmit two digits on the channel for each digit generated by the experimenter. A little thought shows that we can gain nothing by sending a pair of digits each time a coin toss is made. Instead, let us break up the experimenter's digits into a blocks of length two. We have time to replace each block of length two by a block of lenght four. We do so by the encoding dictionary shown in Fig. 4.

0 0 → 0 0 0 0
1 0 → 1 0 0 1
0 1 → 0 1 1 1
1 1 → 1 1 1 0

Fig. 4.

The left side lists the four possible blocks; the right side the corresponding sequence of four digits that is transmitted. At the receiver, the messages are broken up into blocks of four digits. Sixteen different blocks are possible. They all appear on the left side of the decoding dictionary shown in Fig. 5.

0 0 0 0	1 0 0 0	0 1 0 0	0 0 1 0	→	0 0
1 0 0 1	0 0 0 1	1 1 0 1	1 0 1 1	→	1 0
0 1 1 1	1 1 1 1	0 0 1 1	0 1 0 1	→	0 1
1 1 1 0	0 1 1 0	1 0 1 0	1 1 0 0	→	1 1

Fig. 5.

The right side of the dictionary shows the corresponding decoded message that is passed on to the destination. For example, if the received block is either 0000, 1000, 0100, or 0010, then 00 is passed on to the destination. Study of the first row of this table shows that if 0000 is transmitted it is decoded exactly into 00 not only when no errors occur in transmission but also when there is a single error in the first, or the second, or the third digit transmitted. Precisely the same statement holds for each of the three other allowed transmitted blocks of digits, 1001, 0111 and 1110.

The probability of error in digits given to the destination can be computed to be $p = 1/8000$.

We can reduce the error probability further while still permitting the experimenter to toss his coin 500 times per second by an obvious generalization. We break the experimenter's messages up into blocks of length three. There are eight such possible blocks. We prepare a dictionary or code book that replaces each of these eight blocks by a suitably chosen sequence of six binary digits to be transmitted over the channel. At the receiving end, we break the messages up into blocks of length six. There are $64 = 2^6$ possible blocks that might occur. We prepare a decoding dictionary that tells how to replace each such block of six digits by a block of three digits. These are passed on to the destination. With the best such encoding and decoding dictionary, it can be shown that the error probability is now reduced to 10^{-6}.

This procedure can be carried on indefinitely. In the general case, the message is broken into blocks of length k. Each of these 2^k blocks is replaced by a block of n binary digits for transmission over the channel. The received message is broken into blocks of length n. A code book tells how to decode each of these 2^n sequences into an appropriate sequence of k digits. The fundamental theorem and its proof show that by making k and n sufficiently large and by using appropriate encoding and decoding dictionaries of this sort, the probability of error can be made as small as desidered provided only that $k/n < C$.

We do not know how to explicitly construct the dictionaries mentioned above. For a given k and n there will clearly be a minimum value of error probability obtained. We do not know this value except for very small values of k and n. Upper and lower bounds for this best error probability that are asymptotic results for large k and n are known, however. These results show that the error probability can be made to decrease exponentially with n for fixed k/n. From these bounds, it would appear that codes with values of n from 50 to 100 would be extremely useful in application.

However, the code book method described here is clearly out of the question for such values of n, since it must list 2^n entries. If, indeed, such codes are ever to be used in practice, they must have special features, perhaps an algebraic structure, which permit coding and decoding by some calculation technique rather than by dictionary.

The later portion of my talk will be devoted to a discussion of some of the codes that are known. Before proceeding to these however, it seems best to pause in order to prove the fundamental theorem. The proof to follow is an adaptation by E. N. GILBERT to the binary symmetric channel of a more general proof due to A. FEINSTEIN. The proof will, I hope shed some light on the nature of the coding problem.

We shall need a few mathematical results of a secondary nature which we now establish as preliminaries in order not to break the chain of argument at a later point. Suppose we have m distinguishable objects and m different colors of paint. The number of different ways in which we can paint the objects is m^m. This is certainly greater than the number of ways in which we can color the objects when we are restricted to use the first k paints on only k of the objects. This latter number is $\binom{m}{k} k^k (m-k)^{m-k}$, for we can choose the k objects in $\binom{m}{k}$ ways; once chosen they can be colored with the first k paints in k^k ways. The remaining $m-k$ objects can be colored with the remaining $m-k$ colors in $(m-k)^{m-k}$ ways.

Thus

$$\binom{m}{k} \leqslant \frac{m^m}{k^k(m-k)^{m-k}} \,. \tag{1}$$

The next result needed concerns the number w of ones to be found in a sequence of n binary digits when successive digits are produced independently with probability p for a one and probability $q = 1 - p$ for a zero. The probability of finding exactly k ones in such a string is $\Pr(w = k) = \binom{n}{k} p^k q^{n-k}$.

It is easy then to show that $Ew = np$ and that $E(w - np)^2 = np(1 - p)$. Let us now define the number b by

$$b = +\sqrt{\frac{2np(1-p)}{\varepsilon}}\,, \tag{2}$$

where ε is a small number given in advance.

From $E(w - np)^2 = np(1 - p)$ it follows by definition that

$$\sum_{k=0}^{n} \Pr(w = k)(k - np)^2 = np(1 - p)\,,$$

so that

$$\sum_{k>np+b}^{n} \Pr(w = k)(k - np)^2 \leqslant np(1 - p)$$

or

$$b^2 \sum_{k>np+b} \Pr(w = k) \leqslant np(1 - p)\,.$$

The sum, however, is the probability that w be greater than $np + b$.

Combining this result with (2) gives

$$\Pr(w > np + b) \leqslant \frac{\varepsilon}{2}\,. \tag{3}$$

Those familiar with the Chebychev inequality will recognize this as a special case.

A few definitions and another inequality will complete our preliminaries. We shall be concerned with the 2^n n-place binary sequences. We shall refer to the sequences as points and will denote different sequences or points by letters such as x_1, x_2, etc.

We define the distance between two sequences to be the number of places in which they differ, so that, for example, the distance between 1 1 0 1 0 and 0 1 1 0 0 is three since these sequences differ in the first, third and fourth places. By a sphere of radius r about the point x we shall mean the set of all points distant r or less from x.

We shall use the symbol $A(x)$ to denote the sphere of radius $np+b$ about x where b is given by (2). The number of points in such a sphere is

$$N_A = \sum_{k=0}^{np+b} \binom{n}{k}, \tag{4}$$

since there are exactly $\binom{n}{k}$ points distant k from any given point. When n is large and $p < \frac{1}{2}$, $np+b < n/2$ so that the largest term in (4) is the one for $k = np+b$. Thus

$$N_A \leqslant \frac{n}{2}\binom{n}{np+b} \leqslant \frac{n}{2}\,\frac{n^n}{(np+b)^{np+b}(nq-b)^{nq-b}},$$

on using (1), or what is the same

$$N_A \leqslant \frac{n}{2}\,\frac{1}{(p+(b/n))^{np+b}(q-(b/n)^{nq-b}}. \tag{5}$$

We are now ready to prove the fundamental theorem for the binary symmetric channel. We consider transmitting sequences of length n over the channel. We shall choose a particular subset of the 2^n points $x_1, x_2, \ldots, x_{2^n}$, to signal with. We denote these special points by $X_1, X_2, \ldots, X_K$, and call them code points.

We shall also choose K disjoint sets of points $R_1, R_2, \ldots, R_K$, which will define our method of decoding. If a received sequence of n digits is in R_j, we shall assert that X_j was transmitted, $j = 1, 2, ,.., K$.

The sets $R_1, R_2, \ldots, R_K$, are called detection regions.

When a received point does not lie in any detection region we make no decision and count this as an error.

We now proceed to choose the code points and detection regions. X_1 is chosen arbitrarily. R_1 is taken to be the sphere $A(X_1)$. When X_1, is transmitted, the received sequence will lie outside R_1 only if $np+b$ or more errors have occured in transmission. The errors that occur in transmission can be represented by a string of ones and zeros. The ones and zeros representing the errors occur independently with respective probabilities p and $q = 1-p$. By (3) then the probability that the received message lie outside R_1 when X_1 is transmitted is less than $\varepsilon/2$ so that the probability that X_1 be decoded in error $\leqslant \varepsilon/2 < \varepsilon$.

We now proceed to choose X_2. The region R_2 will be taken as that part of $A(X_2)$ that is disjoint from R_1: We choose for X_2 any point such that when transmitted the probability that the received point lie in R_2 be greater than $1-\varepsilon$. The probability that X_2 be decoded in error is less than ε. Thus we have allowed a possible small overlap of the spheres $A(X_2)$ and $A(X_1)$. The probability that the received point lie in this overlap region when X_2 is transmitted must be less than $\varepsilon/2$.

We proceed in this manner selecting successive code points and detection regions. The region R_i is the portion of $A(X_i)$ that is disjoint from $R_1, R_2, \ldots, R_{i-1}$. The point X_i must be chosen so that when X_i is transmitted the probability that the received point lie in R_i be greater than $1-\varepsilon$. We continue in this manner until no new points can be found which qualify as code points. We suppose that a total of K points have been found. From the method by which the code points were chosen and the detection regions formed, it is clear that the probability that any transmitted point be decoded in error is less than ε. The burden of proving the fundamental theorem, then, will be in showing that K is sufficiently large. To this we now turn our attention.

Let R be the union of the detection regions and N_R the number of points in R. Since R is composed of K possibly overlapping spheres each containing N_A points, $N_R \leqslant KN_A$ or

$$K \geqslant \frac{N_R}{N_A}\,. \tag{6}$$

From (5) we have a bound on N_A in terms of the channel and code parameter to continue the inequality (6). There only remains the task of finding a bound on N_R.

To do so, consider the experiment of picking any sequence x at random (*i.e.* each sequence has probability 2^{-n} of being chosen) and transmitting it over the channel. Let y be the received sequence. We observe that

$$\Pr(y\varepsilon R) = \sum_{\text{all } x} \Pr(x \text{ sent}) \Pr(y\varepsilon R/x \text{ sent})\,. \tag{7}$$

Now $\Pr y\varepsilon\,(R/x \text{ sent}) \geqslant \varepsilon/2$ for all x. It is certainly true for the code points $X_1, X_2, \ldots, X_K$. If it were true for some other point x that $\Pr(y\varepsilon R/x \text{ sent}) < \varepsilon/2$, then the probability that y be contained in the overlap of $A(x)$ and R when x is sent would also be $< \varepsilon/2$. The point x would thus qualify as a $(K+1)$-st code point contrary to assumption.

Now $\Pr(x \text{ sent}) = 2^{-n}$ so that (7) becomes

$$\Pr(\text{received point be in } R) \geqslant \frac{\varepsilon}{2}\,. \tag{8}$$

We also have

$$\frac{N_R}{2^n} = \Pr(x\varepsilon R) = \sum_{\substack{x\varepsilon R\\ \text{all } y}} \Pr(x \text{ sent, } y \text{ received})\,. \tag{9}$$

But

$$\Pr\,(x \text{ sent},\ y \text{ received}) = \frac{1}{2^n}\Pr\,(y \text{ received}/x \text{ sent}) =$$

$$= \frac{1}{2^n}\Pr\,(x \text{ received}/y \text{ sent}) = \Pr\,(y \text{ sent},\ x \text{ received})$$

since the error pattern which changes x to y is the same as the error pattern which changes y to x. Thus from (9)

$$\frac{N_R}{2^n} = \sum_{\substack{x\varepsilon R\\ \text{all } y}} \Pr\,(y \text{ sent},\ x \text{ received}) = \Pr\,(\text{received point be in } R)\,.$$

Combination of this result with (8) yields

$$N_R \geqslant \frac{\varepsilon}{2}\,2^n\,.$$

From (5) and (6) then

$$K \geqslant \frac{\varepsilon}{n}\,2^n\left(p + \frac{b}{n^{-}}\right)^{np+b}\left(q - \frac{b}{n}\right)^{nq-b}.$$

On taking the log and dividing by n, one finds

$$\text{(10)} \qquad \frac{\log K}{n} \geqslant 1 + \log q + q \log q - 0(1/\sqrt{n})$$

where the terms indicated by $0(1/\sqrt{n})$ all vanish for large n at least as fast as $1/\sqrt{n}$. The left side of (10) is the rate of transmission.

Thus by making n sufficiently large we can trasmit at a rate as close to the channel capacity $C = 1 + p \log p + q \log q$ as desired with a probability of error less than ε for each code point.

This is the fundamental theorem.

2. – Let us consider now in further detail the matter of constructing codes for the binary symmetric channel. A geometric interpretation of the problem is helpful in understanding it. The 2^n n-place binary sequences can be represented as the vertices of a unit cube in an n-dimensional Euclidean space. The most convenient way to do so is to regard the successive digits of a binary sequence as the successive co-ordinate values that locate the point. Fig. 6

shows the representation of the 8 3-place binary sequences as the vertices of a cube in 3-dimensions.

The selection of a particular code for use on the channel corresponds to specially designating K of the vertices of the cube. When we transmit one of these code points, in general the received sequence is represented by a different vertex of the cube. If a single error occurred in transmission, the received point lies one edge length away from the point sent. If two errors occur, the received point lies two edges away; if j errors occur it lies j edges away. Now the probability of j errors is $\binom{n}{j} p^j q^{n-j}$ which for $p < \frac{1}{2}$ is monotonic decreasing with increasing j. Thus the received point is most likely to be near the transmitted point: less likely to be far away. This comment tells us how to design the best detector for any given code. With each code point, X_i, is associated a region, R_i, such that every point in R_i is at least as close to X_i as it is to any other code point. Decoding is performed by asserting that the transmitted message was X_i whenever the received point lies in R_i. It is not difficult to show that this choice of the detection region R_i leads to minimum probability of error for the given code. Such a decoding scheme is called « a maximum likelihood detector ».

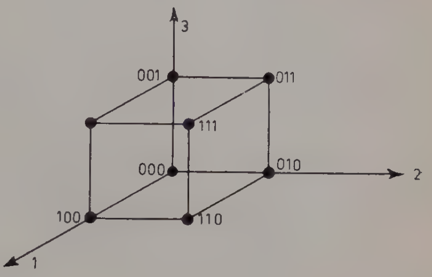


Fig. 6.

In principle, then, the design of the decoding dictionary is clear. The problem, then, is how to designate the special vertices that are to serve as the code. The number of these points is determined by the rate at which we wish to transmit. Their disposition on the cube will determine the error probability to be achieved with the code.

From the geometrical picture, it is easy to see that in some not too precise sense the points of a good code are as far apart from each other as possible. If each of the regions R_i contains the sphere of radius e about X_i, then clearly the code will correct all single, double, ..., e-tuple errors that may occur in the transmission of any code point. Such a code is called an e-error-correcting code.

How many code points can an e-error-correcting code on the n-cube have? The answer is not known. Many upper and lower bounds have been given in the literature, however. We mention only one,

$$\frac{2^n}{\sum\limits_{j=0}^{2e}\binom{n}{j}} \leqslant K \leqslant \frac{2^n}{\sum\limits_{j=0}^{e}\binom{n}{j}} .$$

The right side of this inequality is particularly easy to establish. Each sphere of radius e contains $\sum_{j=0}^{e}\binom{n}{j}$ points. The K spheres may not exhaust all points of the cube, so that $K\sum_{j=0}^{e}\binom{n}{j}\leqslant 2^n$.

It may happen for certain values of n and e that it is possible to find a set of K disjoint spheres of radius e that do exhaust the cube. Such codes are called « close packed e-error correcting codes ». They have some very desirable properties. Unfortunately, there are not many such codes. A necessary condition for the existence of such a code is clearly that $\sum_{j=0}^{e}\binom{n}{j}$ be a power of 2. For example, if $e=1$, we require $\binom{n}{0}+\binom{n}{1}=1+n=2^t$, say, or $n=2^t-1$. This necessary condition turns out to be sufficient in this case, and for $n=2^t-1$, $t=1, 2, \ldots$ there do exist close packed single error correcting codes. These are the well known Hamming codes.

For $e=2$ one finds $\sum_{j=0}^{e}\binom{n}{j}$ a power of two for $n=2$, 5 and 90. It has been shown by H. S. SHAPIRO that these are the only values of n for which this is true. For $n=2$ and 5 quite trivial codes result. LLOYD has shown that for $n=90$, no close packed 2-error correcting code exists.

And so it goes. One can investigate the close packed codes one by one and obtain interesting number theoretic properties, but as a class of codes they do not appear to be too useful. One general theorem due to H. S. SHAPIRO is worth noting before leaving this subject. For any fixed $e\geqslant 2$, there are only finitely many values of n for which $\sum_{j=0}^{e}\binom{n}{j}$ is a power of 2.

Let us turn now to another class of binary codes which has received some attention. These are called group codes or parity check codes. Before describing them, let me remind you briefly of the mathematical notion of a group.

Let I, A, B, C, ... be a distinguishable set of objects (called elements). Let there be given further a law which associates one of the objects with each ordered pair of the objects. This rule is usually written as multiplication, so that, for example, if C is associated with the pair A, B in that order one writes $AB=C$. If the collection of objects and the rule have the following properties, they are called a group:

1) There is a unique element I such that for every element A of the collection $IA=AI=A$.

2) For every element A of the collection there is a unique element called the inverse of A and denoted A^{-1} such that $AA^{-1}=A^{-1}A=I$.

3) For all elements A, B, C

$$A(BC) = (AB)C\,.$$

These are not a minimal set of postulates for a group. Some of the statements are derivable from others. Typical examples of groups are: 1) the set of all positive and negative integers and zero (here the law of association is ordinarily addition); 2) the set of all non-singular $n \times n$ matrices. (Here the group multiplication is ordinary matrix multiplication).

We note that the set of all n-place binary sequences form a group when the group multiplication law is addition modulo 2 of the sequences term by term. Thus if $A = 10111$, $B = 01101$, AB means $10111 + 01101 = 11010$. The identity element of the group is the all zero sequence. Each element is its own inverse. We denote the group by B_n.

A group code is any set of n-place binary sequences that form a subgroup of B_n, *i.e.*, that are also a group under the same law of multiplication. For example, 0000, 1001, 0111, 1110 is a group code. The modulo two sum of any two of these sequences is again in the collection of sequences.

Group codes have many special properties of interest. I can describe only a few of them in the short time remaining.

In the first place, the maximum likelihood regions R_i can be described in a simple manner for these codes. Let us list the elements of the code in a row (first row of Fig. 7). Here I stands for the all zero sequence, and each of the symbols A_2, A_3, ... stands for a particular n-place binary sequence. The elements of this row form a group, so that the product of any two A's is again an A. Let us now define the weight of a binary sequence to be the number of 1's in the sequence. The first row of Fig. 7 does not exhaust B_n. Of the elements in B_n not in the code, choose one of minimal weight and call it S_2. Form the second row of the table as indicated in Fig. 7.

$$\begin{array}{ccccc}
I & A_2 & A_3 & \cdots & A_\mu \\
S_2 & S_2A_2 & S_2A_3 & \cdots & S_2A_\mu \\
S_3 & S_3A_2 & S_3A_3 & \cdots & S_3A_\mu \\
\cdot & & & & \\
\cdot & & & & \\
\cdot & & & & \\
S_\nu & S_\nu A_2 & S_\nu A_3 & \cdots & S_\nu A_\mu
\end{array}$$

$$\mu = 2^K, \qquad \nu = 2^{n-K}$$

Fig. 7.

All of the elements in these two rows are distinct. (For $S_2A_i = S_2A_j$ implies on multiplying by S_2, $A_i = A_j$. Also $S_2A_i = A_j$ implies on multiplying by A_i, $S_2 = A_jA_i$. But S_2 was assumed not in the first row.) Now of the elements in B_n not in the first two rows, choose one of minimal weight, call it S_3 and form the third row. Again it is easily seen that the elements listed are all distinct. We continue in this way until B_n is exhausted. *I* assert that the columns of the array thus constructed are the maximum likelihood regions R_i. That is, the region associated with A_i is A_i, S_2A_i, S_3A_i, ..., $S_\nu A_i$.

To prove this assertion, we must show that any element in the i-th column is at least as close to A_i as it is to any other A. What we have given is that each S_j is of minimal weight in its row. We need therefore first to connect weight with distance. Let $d(R, S)$ denote the distance between elements R and S; let $w(R)$ denote the weight of element R. Then clearly

$$d(R, S) = w(RS) .$$

From this it follows that

$$d(R, S) = d(RT, ST) , \tag{1}$$

for any T, since

$$d(R, S) = w(RS) , \quad d(RT, ST) = w(RTST) = w(RST^2) = w(RS)$$

since $T^2 = I$. Also we note

$$w(T) = d(T, I) . \tag{2}$$

By construction of the array of Fig. 7 $w(S_i) \leqslant w(S_iA_j)$ or from (2)

$$d(S_i, I) \leqslant d(S_iA_j, I) \qquad \text{all } i \text{ and } j.$$

By using (1), this becomes

$$d(S_iA_m, A_m) \leqslant d(S_iA_jA_jA_m, A_jA_m) = d(S_iA_m, A_l) \qquad \text{all } i,\ j,\ m,$$

where we have set $A_jA_m = A_l$. For a fixed m, as j takes all values, so does, l, so we have

$$d(S_iA_m, A_m) \leqslant d(S_iA_m, A_l) \qquad \text{all } i,\ l,\ m.$$

But this asserts that every element in column m is at least as close to A_m as to any other A. Q.E.D.

If w is the weight of an element of B_n, we associate a probability p^wq^{n-w} with the element. Let Q, be the sum of the probabilities associated with the elements of the first column of Fig. 7. Then it is easy to show that Q is the probability that any transmitted A be decoded correctly.

It is convenient to define two group codes to be equivalent if one can be obtained from the other by the application of a fixed permutation of the digits to all the elements of one of the codes. *E.g.* 0000, 1001, 0111, 1110 is equivalent to 0000, 0110, 1011, 1101. The second code is obtained from the first by interchanging the first two digits and by interchanging the last two digits in each code point.

It can be shown that every group code is equivalent to a parity check code. The latter class of codes can be characterized as follows. Let $x_1, x_2, \ldots, x_n$ be the successive digits in a code point: each x is zero or 1. In a parity check code, the digits of every code point satisfy a set of linear (mod 2) equations,

$$x_i = \sum_{j=1}^{k} a_{ij} x_j , \qquad i = k+1, \ldots, n , \qquad a_{ij} = 0 \text{ or } 1$$

and all elements of B_n that satisfy these relations are code points. There are 2^k code points. The first k digits can have any value and are called information digits. The remaining $n-k$ digits are fixed linear mod 2 combinations of the information digits. They are called check digits. The encoding of messages using group codes is thus arithmetic in nature. No dictionary is needed. The incoming message is broken up into blocks of length k to be used as information digits. The $n-k$ check digits are computed and adjoined to the information digits to form the block of n digits to be transmitted.

A simplification also results in the decoding dictionary of a group code. For any received sequence of n digits, form the sequence of $n-k$ digits $r_{k+1} r_{k+2} \ldots r_n$ where

$$r_i = x_i + \sum_{j=1}^{k} a_{ij} x_j , \qquad i = k+1, \ldots, n , \qquad \text{(mod 2)}.$$

This sequence is called the parity check sequence. It can be shown that all elements in any row of Fig. 7 have the same parity check sequence, and that no two rows have the same parity check sequence. To decode, then, one forms the parity check sequence for the received element T. This identifies one of the S's. The product ST then gives the maximum likelihood estimate of the transmitted code point.

Many of the notions just discussed are illustrated in Fig. 8. This will be recognized as the code used as an illustration in Sect. 1. The column at the extreme right lists the parity check sequence for each of the rows.

0000	1001	0111	1110	—	00
1000	0001	1111	0110	—	01
0100	1101	0011	1010	—	11
0010	1011	0101	1100	—	10

$$x_3 = x_2 \qquad r_3 = x_3 + x_2$$

$$x_4 = x_1 + x_2 \qquad r_4 = x_4 + x_1 + x_2$$

Fig. 8.

The problem of finding a group code with maximum Q for given n and k remains unsolved.

3. – There have been many other special codes investigated both for the binary symmetric channel and for other more complex channels. Time limitations prevent discussion of them here. Some of the codes might be useful in certain practical cases, but nothing like a general theory that leads in a constructive manner to the results promised by the fundamental theorem has emerged. The problem remains one of the most challenging in information theory. We have been promised the existence of communication systems with certain highly desirable properties. We do not yet know how to find them, nor do we yet know what price in complexity of equipment must be paid for this promised accuracy of transmission.

The theory of communication must certainly be considered incomplete until answers to these questions have been found.

BIBLIOGRAPHY

R. W. Hamming: *Bell Syst. Techn. Journ.*, **29**, 147 (1950).

E. N. Gilbert: *Bell Syst. Techn. Journ.*, **31**, 504 (1952).

D. Slepian: *Bell. Syst. Techn. Journ.*, **35**, 202 (1956).

S. P. Lloyd: *Bell Syst. Techn. Journ.*, **36**, 517 (1957).

H. S. Shapiro and D. L. Slotnick: *On the Mathematical Theory of Error Correcting Codes* (New York).

General: See the excellent bibliographies of P. E. Stumpers: *IRE Trans.*, Vol. PGIT-2 (1953), Vol. IT-2 (1955), Vol. IT-3 No. 2 (1957).

References to the Russian literature can be found in F. L. H. M. Green: *IRE Wescon Convention Record*, Part **2**, 67 (1957).

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3° Trimestre

A Linear Circuit Viewpoint on Error-Correcting Codes (*).

D. A. HUFFMAN

Massachusetts Institute of Technology - Cambridge, Mass.

1. – Algebraic description and realization of linear sequence filters.

A linear binary sequence filter [1] is a synchronous filter whose inputs and outputs are ordered sequences of binary symbols (0's and 1's). For the general non-time-varying filter each digit of the filter output sequence is a modulo-two sum of an arbitrary selection of past output digits (Z) and of present and past input digits (X). The description of a sequence filter in terms of a delay operator, D, is a straightforward one. For example, a filter whose output Z is the sum of the first and third previous output digits and of the present, first, second, and fourth previous input digits is described by

$$Z = DZ + D^3Z + X + DX + D^2X + D^4X\,, \tag{1}$$

where the + symbol is used here for the modulo-two operation. That is, the present output is zero if an even number of selected digits have the value one, and is unity if an odd number have the value one.

Since the modulo-two operation is self-inverse, the terms in (1) may be rearranged to give

$$D^3Z + DZ + Z = D^4X + D^2X + DX + X \tag{2-a}$$

or

$$(D^3 + D + I)Z = (D^4 + D^2 + D + I)X\,. \tag{2-b}$$

(*) Excerpts from *IRE Transactions on Information Theory*, Vol. IT-2, (Sept. 1956 T).

The « transfer ratio » of the filter is then

$$\frac{Z}{X} = \frac{D^4 + D^2 + D + I}{D^3 + D + I}\,. \tag{3}$$

An efficient realization of this filter results from rearranging (1) to give:

$$X + Z = D(X + Z) + D^2X + D^3Z + D^4X \tag{4-a}$$

or

$$X + Z = D\big\{(X + Z) + D\{X + D\{Z + DX\}\}\big\}. \tag{4-b}$$

The corresponding filter is given in Fig. 1-*a*. The « inverse » filter, whose input is Z and whose output is X is described by Eq. (4-*a*, *b*) and has a transfer ratio

$$\frac{X}{Z} = \frac{D^3 + D + I}{D^4 + D^2 + D + I}\,. \tag{5}$$

Its realization is given in Fig. 1-*b*. Both of the filters in Fig. 1 utilize only two kinds of elements: modulo-two adders and unit delays (single-stage shift-registers). The « chain » realization given both of these filters consists of a chain of unit delays with provision made for introducing the signals X, Z or $(X + Z)$ between each two stages of delay. It uses just the number of delay units necessary to remember the input or output digit most remote in the past which is needed for proper operation of the filter (in this case the fourth previous input), and an equal number of adders.

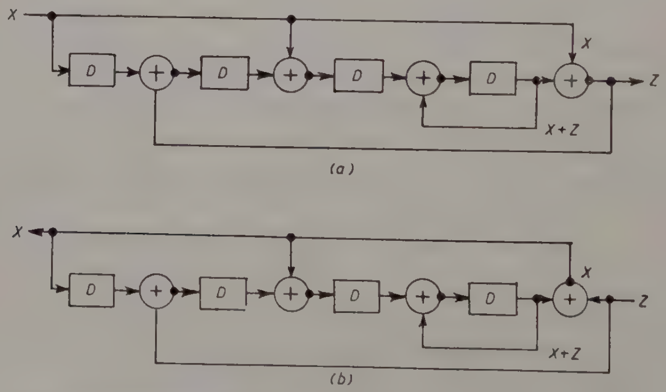


Fig. 1. – Chain realization of a binary sequence filter and its inverse.

When a binary filter and its inverse are connected in cascade, one mode of operation of the combination is that for which the transfer ratio is the identity operator. In our example:

$$\left(\frac{D^4+D^2+D+I}{D^3+D+I}\right)\cdot\left(\frac{D^3+D+I}{D^4+D^2+D+I}\right)=I\,. \tag{6}$$

That is, the second filter unscrambles the scrambling produced by the first. In the error-correcting scheme proposed in this paper the use of filters and their inverses will be of paramount importance.

2. – Description of a filter from its impulse response characteristics.

Later in this paper we will want to make use of the fact that there exist finite realizations of linear sequence filters whose response to an input « impulse » (a single digit 1 preceded and followed by infinite sequences of 0's) is arbitrary except that it must eventually die out (become the all-0 sequence) or ultimately become periodic. Suppose, for example, that we wish to realize a filter whose response to an input sequence, X^*, containing an impulse is the output sequence, Z^*, which ultimately becomes periodic (see Fig. 2a).

X^*: . . . 0 0 0.1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 . . .

Z^*: . . . 0 0 0.1 0 1 0 1 0 0 1 1 1 0 1 0 0 1 1 1 0 1 0 0 . . .

Z_p^*: . . . 0 0 0.1 1 1 0 1 0 0,1 1 1 0 1 0 0,1 1 1 0 1 0 0, . . .

Z_t^*: . . . 0 0 0.0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0. . . .

(a)

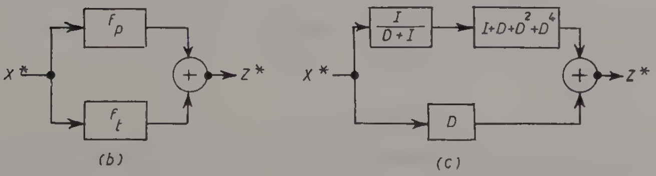


Fig. 2. – Steps in the synthesis of a binary filter form a specified impulse response.

Z^* can always be considered to be the sum of two sequences: Z_p^*, the periodic component, and Z_t^* the transient component. The filter we are trying to design may, for the moment, be considered to be made up of two sub-filters, f_p and f_t, which have impulse responses Z_p^* and Z_t^*, respectively, and whose outputs are added to give the desired response Z^* (see Fig. 2-b).

The filter f_p could be realized by a cascade of two other filters (see Fig. 2-c). The first would have an impulse response which consisted of a sequence of impulses spaced seven intervals apart (the period of the periodic response) and continuing indefinitely. This filter would have a transfer ratio

$$I + D^7 + D^{14} + D^{21} + \dots = \frac{I}{D^7 + I}\,. \tag{7}$$

The periodically recurring output of this filter could be used as the input to another filter having the proper transient response (finite in length). The latter filter has a transfer ratio which is a polynomial, $I+D+D^2+D^4$ in this example, whose terms correspond to the positions of the 1's in a typical cycle, contained between commas, of the desired Z_p^*.

The transient part, Z_t^*, of the impulse response is easy to arrange for in our example. The proper associated filter, f_t, has a transfer ratio D.

The filter we are designing could then be realized with a total transfer ratio of

$$\frac{Z^*}{X^*} = \left[\frac{I}{D^7 + I}\right](I + D + D^2 + D^4) + D\,, \tag{8a}$$

which may be rewritten as

$$\frac{Z^*}{X^*} = \frac{(I + D + D^2 + D^4) + D(D^7 + I)}{D^7 + I} \tag{8b}$$

or as

$$\frac{Z^*}{X^*} = \frac{D^8 + D^4 + D^2 + I}{D^7 + I}\,. \tag{8c}$$

The numerator and denominator of the expression in Eq. (8-c) each contain the factor $D^4 + D^2 + D + I$ (found using the Euclidean algorithm; see reference [1]) which may be cancelled to give

$$\frac{Z^*}{X^*} = \frac{(D^4 + D^2 + D + I)(D^4 + D^2 + D + I)}{(D^4 + D^2 + D + I)(D^3 + D + I)} = \frac{D^4 + D^2 + D + I}{D^3 + D + I}\,. \tag{8d}$$

The transfer ratio is now in its simplest form and the filter may be synthesized as has already been done in Eqs. (4) and Fig. 1.

3. – A linear single-error correcting coding scheme.

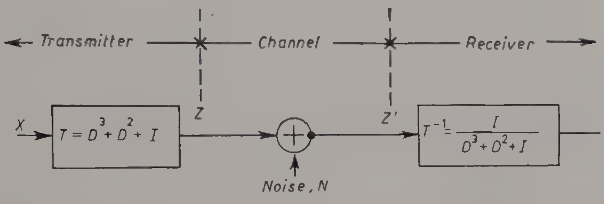
Consider the arrangement of filters shown in Fig. 3-a. A sequence of seven X digits is fed into a transmitter filter with transfer ratio T, resulting in a sequence $Z = (T)X$ which is transmitted through the noisy channel. In the channel a noise sequence, N, is added to Z so that what arrives at the receiver filter is

$$Z' = Z + N\,. \tag{9-a}$$

At the receiver a filter inverse to the transmitter filter creates from the sequence Z' a sequence

$$X' = (T^{-1})Z' = (T^{-1})(Z+N) = (T^{-1})\left[(T)X + N\right] = X + (T^{-1})N\,. \tag{9-b}$$

If there were no noise in the channel ($N = 0$), $X = X'$. If there is noise present then the sequence X' contains the sum of the transmitter input sequence, X and the response $(T^{-1})N$ of the receiver filter to the noise. If only a single noise digit is present, the sequence X' contains X plus the impulse response of the receiver filter superimposed thereon.



(a)

X: (1 1 1 0) 0 0 0 $\qquad\qquad\qquad\qquad$ Z': 1 1 1 0 0 1 0

N: 0 0 1 0 0 0 0

Z: 1 1 0 0 0 1 0 $\qquad\qquad\qquad\qquad$ X': 1 1 0 0 1 1 1

(b)

$$\begin{array}{rl} \text{X:} & (1\,1\,1\,0)\,0\,0\,0 \\ (T^{-1})\ N: & 0\,0\,1\,0\ \ 1\,1\,1 \\ \hline X' = X + (T^{-1})\ N: & 1\,1\,0\,0\ \ 1\,1\,1 \end{array}$$

(c)

Impulse response of the receiver filter with transfer ratio.

$$T^{-1} = (D^3 + D^2 + I)^{-1}:$$

$$\ldots 0\,0\,0\,0\,0.1\,0\,\underbrace{1\,1\,1}\,0\,0,1\,0\,1\,1\,1\,0\,0,1\,0\,1\ldots$$

(d)

Fig. 3. – An elementary example of the linear single-error detecting scheme.

Let us examine the coding and decoding mechanism in more detail. The first four digits of the sequence X are information digits, and may therefore be chosen in $2^4 = 16$ different ways. The remaining three digits are always all zeros and are to be called here buffer digits. The composite block of seven digits is scrambled for transmission in the channel by the first filter. The sequence X' which results from the unscrambling action of the receiver filter would equal X if there were no noise in the channel. The clue to this possibility would be the existence of three zeros in the buffer positions (the last three digits) in the sequence X'.

When a single noise digit is equal to unity (just one transmitted digit is changed by noise action) the received sequence, X', may look quite different from X (see Fig. 3-*b*). In particular, the buffer positions will no longer contain all zeros, but will instead be three successive digits of the impulse response of the receiver filter. In our example the impulse response of that filter is given in Fig. 3-*d*, and since we have assumed the noise impulse to occur in the third position of the block of seven digits, we observe in the buffer positions the third, fourth, and fifth digits of the impulse response (see Fig. 3-*c*,*d*).

It is extremely important to notice that the digits in the buffer positions of the sequence X' are *independent* of which of the sixteen possible X sequences is sent. This pattern of digits depends only upon the position(s) of the noise digits and upon the impulse response of the receiver filter.

We have chosen the receiver filter so that the impulse response has a period of seven digits (the length of the composite block) and so that each of the seven possible combinations of three (the number of buffer digits and the degree of the denominator polynomial) successive digits in the response will be different from the others. That this is possible for a block length of $n = 2^b - 1$ with b buffer positions follows from the fact that the maximum possible period of the impulse response of a filter with denominator polynomyal of degree b is $2^b - 1$ [1].

By observing the three buffer positions of the sequence X', and by knowing the form of the impulse response of the receiver filter, we can deduce where the noise impulse occurred in the block. If we assume that only a single noise impulse was present (the most likely situation) we can recreate the original sequence X by adding (same as subtracting, modulo-two) the now known sequence $(T^{-1})N$ to the sequence X'.

For our example it is interesting to examine the sixteen possible sequences, Z, which correspond to the sixteen possible sequences, X, which might be inserted into the transmitter filter with $T = D^3 + D^2 + I$. These are listed in Fig. 4. The sixteen Z sequences are mutually separated by a distance of at least three, a necessary condition for single-error correction [2].

The advantage of the linear circuit viewpoint of this paper is that instead of concerning ourselves with the distance properties of $2^k = 2^{n-b}$ (in our example, 16) different code message sequences $Z = (T)X$, we may concentrate our attention on the impulse response of the receiver filter with transfer ratio T^{-1}. It is not claimed that this latter viewpoint will ultimately be more advantageous than the first, but only that two viewpoints are better than one.

X:		$Z = (T)X$:
0 0 0 0 ⋮ 0 0 0	→	0 0 0 0 0 0 0
0 0 0 1 ⋮ 0 0 0		0 0 0 1 0 1 1
0 0 1 0 ⋮ 0 0 0		0 0 1 0 1 1 0
0 0 1 1 ⋮ 0 0 0		0 0 1 1 1 0 1
0 1 0 0 ⋮ 0 0 0		0 1 0 1 1 0 0
0 1 0 1 ⋮ 0 0 0		0 1 0 0 1 1 1
0 1 1 0 ⋮ 0 0 0		0 1 1 1 0 1 0
0 1 1 1 ⋮ 0 0 0		0 1 1 0 0 0 1
1 0 0 0 ⋮ 0 0 0		1 0 1 1 0 0 0
1 0 0 1 ⋮ 0 0 0		1 0 1 0 0 1 1
1 0 1 0 ⋮ 0 0 0		1 0 0 1 1 1 0
1 0 1 1 ⋮ 0 0 0		1 0 0 0 1 0 1
1 1 0 0 ⋮ 0 0 0		1 1 1 0 1 0 0
1 1 0 1 ⋮ 0 0 0		1 1 1 1 1 1 1
1 1 1 0 ⋮ 0 0 0		1 1 0 0 0 1 0
1 1 1 1 ⋮ 0 0 0	→	1 1 0 1 0 0 1

Fig. 4. – Coded sequences for single-error correction ($n = 7$).

For single-error correction in a block of length n containing b buffer positions $k = n - b$ information positions we need only have a receiver filter with an impulse response with period of length n with each b successive digits in that response different from each other subsequence of length b. This is possible for the case $n = 2^b - 1$ and the proper polynomial is one of degree b which has a maximal-length « null sequence » [1] of $2^b - 1$ digits. Several of these are listed in Fig. 5.

$D^2 + D + I$

$D^3 + D^2 + I$

$D^4 + D^3 + I$

$D^5 + D^3 + I$

$D^6 + D^5 + I$

$D^7 + D^6 + I$

Fig. 5. – Polynomials having maximal length null sequences.

4. – Conclusions.

Extensions of the preceding ideas apply to the situation in which more than one error occurs in the channel, and were treated in the paper from which this is an excerpt. The main conclusion reached in that paper was that instead of thinking about the distance properties of 2^k message points in an n-dimensional space, we may profitably think of designing a linear binary sequence filter at the receiver whose impulse response is of such a form that, by viewing $b = n - k$ successive digits of it we distinguish subsequences due to single errors, by viewing b digits of two superimposed impulse responses we may distinguish sub-sequences due to double errors, etc.

REFERENCES

[1] D. A. Huffman: *The synthesis of linear sequential coding networks*, in *Proceedings of the third London Symposium on Information Theory* (September, 1954).

[2] R. W. Hamming: *Error detecting and error correcting codes*, in *Bell Syst. Techn. Journ.* (1954), pp. 147-160.

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3º Trimestre

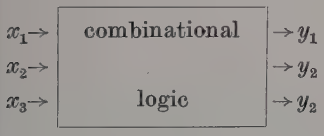
Notes on Information-Lossless Finite-State Automata.

D. A. HUFFMAN

Massachusetts Institute of Technology - Cambridge, Mass.

1. – Combinational circuits.

An information-lossless transducer is roughly one for which a knowledge of the output sequence of symbols is sufficient for the determination of the corresponding input symbols. For a combinational circuit the definition is particularly simple since for such circuits the input symbol (or combination of symbols) at a given moment uniquely defines the output symbol (or combination of symbols). Thus a information-lossless combinational circuit is one for which no two different input combinations can produce the same output combination. For instance, for a combinational circuit with n input and n output leads upon which binary signals can appear the usual truth-table or table of combinations may be examined easily to determine whether or not the requisite one-to-one mapping is represented, and the logical equations expressing the output symbol values in terms of the input symbol values may be solved for the input symbols in terms of the output symbols.



(*a*)

$$\begin{cases} y_1 = 1 + x_1 + x_3 + x_1x_2 + x_1x_3 \\ y_2 = 1 + x_1 + x_2 + x_3 \\ y_3 = 1 + x_1 + x_2 + x_1x_2 + x_2x_3 \end{cases}$$

(*b*)

x_1	x_2	x_3	y_1	y_2	y_3
0	0	0	1	1	1
0	0	1	0	0	1
0	1	0	1	0	0
0	1	1	0	1	1
1	0	0	0	0	0
1	0	1	0	1	0
1	1	0	1	1	0
1	1	1	1	0	1

(*c*)

Fig. 1. – Illustrating a lossless combinational circuit.

As an example consider the circuit described in Fig. 1. The equations of Fig. 1-*b*, written in terms of the operations of the logical product and addition mod-2, are seen for this case to be non-linear since they contain product terms such as x_1x_3.

The rows of entries in the right-hand side of the table of combinations of Fig. 1-*c* are seen to be the same as the rows of entries in the left-hand side, although these rows have been rearranged.

In Fig. 2 are demonstrated the « inverse » table of combinations and solutions for the x's in terms of the y's. A simple test for the solvability of the equations in which the y's are expressed in terms of the x's is that the expressions for y_1, y_2, y_3, y_1y_2, y_1y_3, and y_2y_3 should *not* contain the term $x_1x_2x_3$ but that this term *should* be contained in the expansion of the expression for $y_1y_2y_3$.

y_1	y_2	y_3	x_1	x_2	x_3
0	0	0	1	0	0
0	0	1	0	0	1
0	1	0	1	0	1
0	1	1	0	1	1
1	0	0	0	1	0
1	0	1	1	1	1
1	1	0	1	1	0
1	1	1	0	0	0

(*a*)

$$\left\{\begin{aligned} x_1 &= 1 + y_1 + y_3 + y_1y_2 \\ x_2 &= y_1 + y_2y_3 \\ x_3 &= y_2 + y_3 + y_1y_2 + y_2y_3 \end{aligned}\right.$$

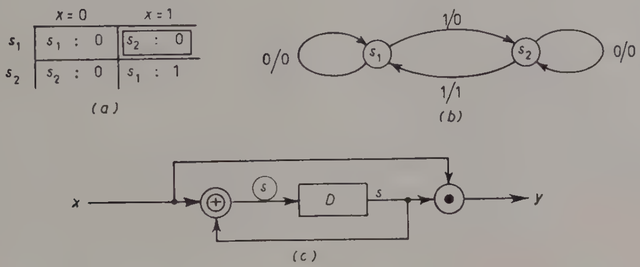
(*b*)

Fig. 2. – The description of a circuit inverse to that of Fig. 1.

2. – Terminal description of sequential circuits.

For sequential circuits a more detailed definition of information-losslessness is necessary. We consider sequential circuits for which a knowledge of the present state and the circuit input determines the next following state and the corresponding output. We will limit ourselves to circuits with a finite number of states, for which the outputs are associated with the transitions between states and occur in synchronism with the corresponding input symbol which produced them, and which have a single binary input and a single binary output. (Extension to circuits with an infinite number of states, or with outputs associated with states rather than transitions, or with input and output symbols chosen from a larger alphabet will not be made here because they would not add to our fundamental understanding.)

All such finite-state automata may be described by any method which shows the dependence of the next state (S) and of the output (y) upon the present state (s) and the input (x). A flow table (as in Fig. 3-a) has the advantage of compactness and orderliness of presentation of the necessary data, and a state diagram (as in Fig. 3-b) perhaps has the advantage of giving a better feel for what sequences of states are possible. The correspondence between these two forms may be illustrated by reference to the upper right entries in the flow table and the heavily lined transition of the flow graph. Each of these is interpreted. When the circuit is in state s_1 and if the input symbol is $x = 1$, the resulting output symbol is $y = 0$ and the next state is s_2. One possible circuit which has the terminal characteristics of Fig. 3-a, b is shown in Fig. 3-c.



	$x = 0$	$x = 1$
s_1	s_1 : 0	$\boxed{s_2 : 0}$
s_2	s_2 : 0	s_1 : 1

Fig. 3. – A sequential circuit description and realization.

The dependence of (s) and y upon s may be displayed in the block diagram form of Fig. 4. The general problem of synthesis (not treated here) is to determine for an arbitrary state diagram or flow table how many feedback loops are necessary and what specific functions should be incorporated in the combinational logic.

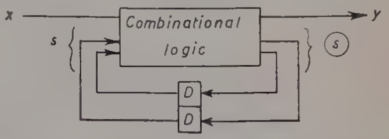


Fig. 4. – General form of a sequential circuit.

3. – Definition of information quantities in sequential circuits.

The information quantities which we are going to use here are related to the knowledge that an observer of the circuit has when he has a knowledge of the describing flow table and of the sequence of output symbols, but no direct knowledge of its input symbols or of its internal states (*).

(*) These quantities are defined more precisely and illustrated more fully in *Information Conservation and Sequence Transducers*, in *Proceedings of the Symposium on Information Networks*, pp. 291-307, Polytechnic Institute of Brooklyn, April, 1954.

Input information is related to the output observer's expectation of a given input symbol. If, for example, the binary input symbols are equally likely and independent of each other the input information rate is at all times one bit per symbol.

Output information is related to the output observer's expectation of a given output symbol. In the circuit of Fig. 3 this observer knows that the state s_1 can be followed only by transitions which yeld the output $y=0$. Therefore when he knows that the state of the circuit is s_1 and observes that the output is $y=0$ the corresponding output information is zero.

Information is stored when, from the output observations only, it becomes impossible to tell exactly what the state of the circuit is. If, for example, an observer calculates (as he might if given the data in the paragraph above) that the circuit is in the state s_1 or state s_2 with equal probability, then for him the circuit has stored one bit of information. Note that the quantity of information stored in this sense may be arbitrarily large even for a circuit with only two states and a correspondingly simple realization if only the input symbols are unexpected enough to the observer.

Information is *lost* when change of internal state takes place so as to eliminate (wholly or partially) data about the past history of the circuit input. Its measure is related to the probability that the actual input symbol sequence was responsible for the observed output sequence rather than any of the other possible input sequences. For example if the output observer knows that the initial state of our circuit is s_1 and then sees two zeros in succession as output symbols then, for him, information is lost even if the final state of the circuit is now revealed to be s_2, since the corresponding input sequence could have been either 0,1 or 1,0 and no further analysis of the output data preceding the initial state of s_1 nor the output data following the final state of s_2 will be of any avail in determining which input sequence actually occurred (see Fig. 5-*a*). If these sequences were equally likely one bit has been lost.

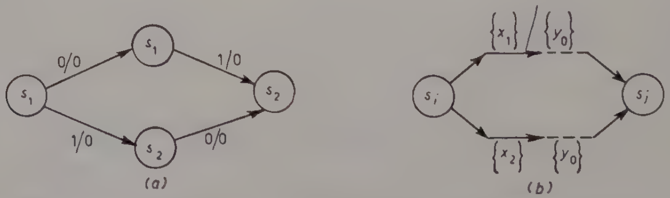


Fig. 5. – Illustration of conditions for information loss.

It may be proved that with those definitions of information quantities the following information conservation is valid for each step in an indefinitely long sequence of observations:

$$I_{\text{input}} = I_{\text{output}} + I_{\text{lost}} + \Delta I_{\text{stored}} .$$

4. – Definition of information-lossless automata.

It is clear from the preceeding discussion that information loss occurs in a circuit when two or more input sequences may lead to the same output sequence because then the input sequence cannot be uniquely determined if only the output sequence is known. More exactly, a sequential circuit (even one with an infinite number of states) is defined as lossless if and only if there exist no two (not necessarily different) states s_1 and s_2, and no two different equal-length input sequences $\{x_1\}$ and $\{x_2\}$ and no output sequence $\{y_0\}$, such that both $\{x_1\}$ and $\{x_2\}$ can lead from s_1 to s_2 and yield $\{y_0\}$. This is of course equivalent to saying that a circuit is lossless if and only if, for an indefinitely long experiment in which the initial and final states and the output sequence are known, the input sequence may be uniquely determined.

5. – Class I information-lossless automata.

The clerical procedures for the determination of losslessness may be organized rather neatly. Consider the flow table of Fig. 6-*a* and the derived table of Fig. 6-*b*. The first row of this derived table tells us that if the initial state of the circuit is s_1 the next state may be deduced to be either s_4 or s_3 immediately upon determination of the output symbol as $y=0$ or $y=1$, respectively. The other rows have similar interpretations. Clearly the example before us is a special case for which an input symbol always produces an immediate (mod-2) effect upon the output and is characterized by the fact that each of the two transitions away from any state are associated with the two different output symbols. Thus the possibility for « parallel » sequences shown in Fig. 5-*b* does not exist. For such circuits, which will be called Class I circuits, it is possible to derive *inverse* circuits which when put in cascade with the original produce as their output sequence an exact replica of the input sequence of the original. The terminal specifications for these inverse circuits are easily

	$x=0$	$x=1$
s_1	$s_3:1$	$s_4:0$
s_2	$s_4:0$	$s_1:1$
s_3	$s_4:1$	$s_2:0$
s_4	$s_3:0$	$s_2:1$

(*a*)

	$y=0$	$y=1$
s_1	$s_4:(1)$	$s_3:(0)$
s_2	$s_4:(0)$	$s_1:(1)$
s_3	$s_2:(1)$	$s_4:(0)$
s_4	$s_3:(0)$	$s_2:(1)$

(*b*)

Fig. 6. – Tabular test for losslessness applied to a Class I circuit.

had by completing the table illustrated in Fig. 6-*b* with entries (paranthesized) telling what x-symbol should be associated with a given transition.

A block diagram showing one possible realization of a Class I circuit is shown in Fig. 7-*a* and one possible realization of its inverse is shown in Fig. 7-*b*. These two circuits differ only in the connections made to the mod-2 adder gate, and therefore we may conclude that the inverse to a Class I circuit may be realized in a circuit having the same number of states as the original. Moreover, since the circuits have a reciprocal relationship either may be used as the canonic form of a Class I circuit.

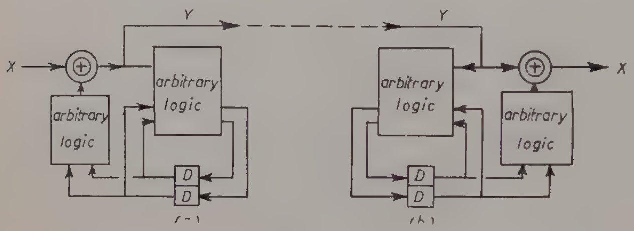


Fig. 7. – Canonic forms for a Class I circuit.

6. – Class II information-lossless automata.

Another case of an information-lossless circuit is shown in Fig. 8. The upper four rows of the table in Fig. 8-*b* are derived in a manner similar to that used for our previous example, except that now the knowledge of an output symbol does not lead immediately to a knowledge of the input symbol which produced it. For example, the second row of the derived table is to be interpreted as follows: If the initial state of the original circuit is s_2 then the symbol $y=0$ must necessarily follow and as a result we are now not certain whether the following state is s_1 or s_3 (or whether the input symbol was $x=0$ or $x=1$).

	$x=0$	$x=1$
s_1	$s_2:0$	$s_3:1$
s_2	$s_1:0$	$s_3:0$
s_3	$s_4:1$	$s_1:1$
s_4	$s_2:1$	$s_4:0$

(*a*)

	$y=0$	$y=1$
s_1	s_2	s_3
s_2	s_{13}	—
s_3	—	s_{14}
s_4	s_4	s_2
s_{13}	s_2	s_{134}
s_{14}	s_{24}	s_{23}
s_{23}	s_{13}	s_{14}
s_{24}	s_{134}	s_2
s_{134}	s_{24}	s_{1234}
s_{1234}	s_{1234}	s_{1234}

(*b*)

Fig. 8. – Tabular test for losslessness applied to a Class II circuit.

The first four rows of the new table indicate that confusion may exist between states s_1 and s_3 or between s_1 and s_4. The two symbols s_{13} and s_{14} are entered as designators for rows which are added to the first four rows of the table. Entries for these new rows are found by adding subscripts found in the corresponding entries found in the rows specified by the subscripts of the designator of the new row. For instance, the entry in the $y = 1$ column for the row headed s_{13} is s_{134} since the entries found in the rows headed s_1 and s_3 were s_3 and s_{14}. The newly derived entry tells us that if we are uncertain as to whether the state of the circuit is s_1 or s_3 and if an output symbol $y = 1$ is observed, our new uncertainty is among s_1, s_3 and s_4. The process of generation of new rows is repeated as long as is necessary. Ultimately the necessity for new rows is ended and the table is complete. If in the process of adding subscripts from « component » rows to find the subscripts for « composite » rows no situation is found in which the same subscript is found in each of the component rows, the circuit being tested is information-lossless. Our present example is one of this type.

It could have been seen directly from Fig. 8-*a* that the flow table described a lossless circuit, since two and only two transitions lead to each state and each of these transitions is associated with a different output symbol. We will call such a circuit a Class II circuit. Thus there is no possibility for « parallel » sequences shown in Fig. 5-*b*. Further, a knowledge of the final state of the circuit and the last output symbol is enough for the determination of the next-to-final circuit state. Thus the input sequence for a finite experiment on a Class II circuit may be determined from a knowledge of the *final* state and the output sequence, just as the input sequence for a finite experiment on a Class I circuit may be determined from a knowledge of the *initial* state and the output sequence.

Since, for a Class II circuit, knowledge of a state and the output symbol for the transition leading to that state is sufficient for the determination of the preceding state and this input symbol, this is equivalent to saying that the combinational logic of the general block diagram of Fig. 4 is, for Class II circuits, lossless in the sense of Section **1** (see Fig. 9).

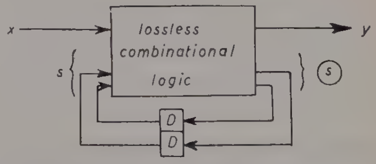


Fig. 9. – Canonic form for a Class II circuit.

7. – More general information-lossless circuits.

It seems to the author that both Class I and Class II circuits deserve to be called information-lossless, the first since an inverse circuit can always be specified and the second both because a specific decoding procedure can be described once the final state of an experiment is given, and because of the conceptually satisfying result that a lossless combinational circuit in which some outputs are reintroduced as inputs after a unit delay is also lossless in the wider sense we have used in this paper to apply to sequential circuits. It is only fair to point out to the reader that some other more restricted definitions of terms similar to information-losslessness as used in this paper have been used and probably will continue to be used by others.

There are many circuits which are lossless which are neither purely Class I or purely Class II circuits. For all of these circuits the test illustrated in Fig. 8-*b* is valid, but the circuit cannot be synthesized in either of the canonic forms already given. Instead the more general canonic form shown in Fig. 10 may be shown (by specifying a rather involved synthesis procedure) to be appropriate. It is easy to see that the canonic forms of Fig. 7-*a* and Fig. 9 are special cases of the diagram in Fig. 10. The meaning of the block labelled « lossless logic » with feedback signals (β) acting as « control » is that, for any set of values of the β-signals, the values of x and α may be determined from a knowledge of the values of y and (α).

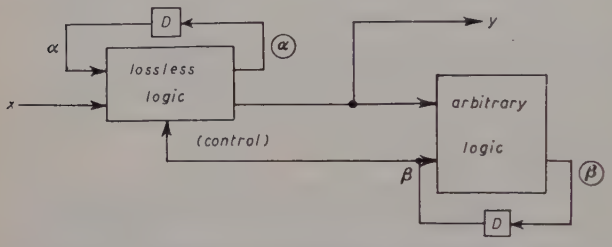


Fig. 10. – General canonic form into which all information-lossless finite automata may be synthesized.

An interesting result which was not apparent until the canonic form of Fig. 10 was derived is that the total internal state at both the beginning and at the end of an experiment need not be known. Instead, information about the initial state of the feedback loops around which the β-signals flow and about the final state of the feedback loops around which the α-signals flow (along with a knowledge of the sequence of y-symbols) is sufficient for determination of the sequence of input symbols. The actual decoding procedure involves the determination of the entire sequence of β-values from the given data about the initial β-value and the sequence of y-values, and next an iterated determination of the sequences of α-values and x-values from the final α-value and the now-known sequence of β-values.

Also of interest is the fact that no information is ever stored in the right-hand portion of the circuit since knowledge about the initial value of the β-signals and about the sequence of y-symbols gives us an always up-to-date exact knowledge of what the signals are in the β-feedback loop(s). All the stired information in the circuit is associated with the block marked « lossless logic » and the adjacent feedback loop(s).

Finally, to serve a warning to those who still somehow feel that « information » is associated with the symbols within a finite-state circuit on a sort of every-symbol-carries-its-own-information basis we show the circuit of Fig. 11, which has no feedback loops, but in which, nevertheless, information necessary for the determination of some input symbol may be stored for an arbitrary large number of steps of an experiment. Of course, then, no finite « inverse » exists even if we agree that the « inverse » need not regenerate the x-sequence immediately, but only after some long but fixed delay. The details of analysis of this circuit are left to the interested reader.

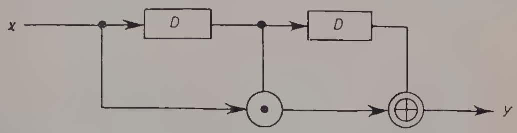


Fig. 11. – An interesting finite state circuit.

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3º Trimestre

Application of Statistical Notions to Multipath Channels.

P. E. GREEN, jr.

Massachusetts Institute of Technology, Lincoln Laboratory - Lexington, Mass.

I should like to discuss some results obtained by my colleagues and myself in trying to devise, by analysis of suitable statistical models, an operating communication system for multipath conditions. This work has already been described in great detail elsewhere (PRICE and GREEN [1]); here I shall summarize it for those to whom these ideas are new.

The problem is somehow to devise a system for sending binary information through a channel perturbed by randomly varying multipath and additive gaussian noise. By multipath we shall mean a condition in which there is more than one propagation path from transmitter to receiver, and we shall begin by postulating a suitable model for this often troublesome phenomenon. For the practical cases of interest we can say that

1) The maximum difference in the travel times of the transmitted signal along the different propagation paths is some number T_M. The strengths of paths outside this range T_M are to be considered negligible.

2) The times-of-flight and strengths of the individual paths are random variables with some upper limit R on fluctuation rate to be defined shortly.

3) The propagating medium is linear.

The particular communication environment that has been of interest in this work has been the so-called high frequency band ((3 ÷ 30) MHz) in which multipath has always been a severe problem. We are interested in sending binary information where a duration T_D is allotted to the transmission of each successive binary digit.

We will be interested in situations in which

$$T_M < T_D < 1/R\,, \tag{1}$$

and the results described here will be applicable to any such situation. In the high frequency problem T_M is usually less than five milliseconds, T_D is several tens of milliseconds and R is of the order of $\frac{1}{3}$ to 3 Hz.

The three assumptions just given allow one to define a time-varying linear filter as a model for the propagating medium (Fig. 1-A). The response to a unit impulse is $h(\tau)$ typically but not necessarily representable as a series of impulses as in Fig. 1-B changing slowly with real time t. Fig. 1-B, shows series $h(\tau, t)$ for a particular value of t. The time-varying complex frequency response $H(\omega)$ at a time t is defined as the Fourier transform of that function $h(\tau)$ occurring at time t. Under a fourth assumption, namely that

4) The communication is confined to a frequency band W cycles per second in width with W smaller than the center frequency of the band.

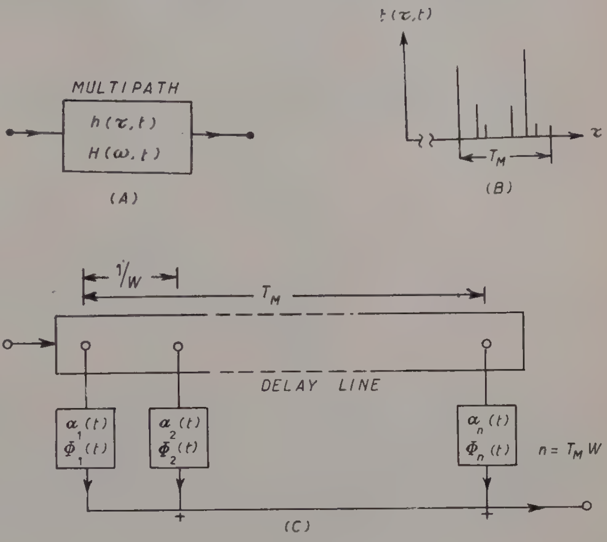


Fig. 1.

The equivalent filter of Fig. 1-A can be redrawn in the form shown in Fig. 1-C. This is done by using a form of the sampling theorem stating that the function $h'(\tau)$, the inverse Fourier transform of that portion of $H(\omega)$ lying in W, can be represented uniquely in terms of $2T_M W$ suitable samples. The particular samples chosen are values of amplitude and phase of $h'(\tau)$ taken at values of delay spaced $1/W$ second apart. These amplitudes and phases α_i and φ_i are slowly varying functions of time t and are represented as the gains and phase shifts of amplifiers whose inputs are the outputs of a tapped delay line, and whose outputs are all added together. We arbitrarily define the parameter R, as the reciprocal of the smallest time during which any α_i of Φ_i changes by say 10 percent or 10 degrees respectively.

Once we have defined what we mean by a multipath channel, and have set down a model to represent it, it is possible to proceed in several directions in deriving an optimum communication system to work through it. The derivation to be presented now leads to a form of optimum receiver which has been called a « rake » receiver. One can derive the same configuration in different ways. However, in each derivation that we have been able to devise there is at least one step that must be made in a somewhat heuristic way. No one analysis is completely self-contained.

To deduce the form of optimum receiver for both additive noise and multipath present it is convenient to start with the well-known optimum receiver for additive noise alone, and then generalize it to include multipath as well.

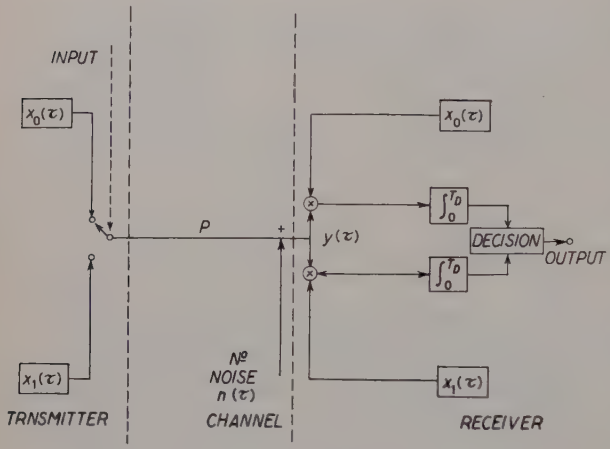


Fig. 2.

Fig. 2 depicts this simpler situation. At the left is a transmitter sending one of two waveforms $x_0(\tau)$ or $x_1(\tau)$ to represent symbols 0 or 1. Each waveform $x_0(\tau)$ and $(x_1)\tau$ is assumed to have negligible energy outside a time interval T_D seconds long. A succession of these signals is sent and in the channels corrupted by the addition of a stationary noise function $n(\tau)$ having a gaussian amplitude distribution and a constant spectral power density N_0 W/Hz over the frequency region of interest. The optimum receiver is defined as one which will deliver at its output a succession of 0 or 1 symbols which differ from the input sequence as infrequently as possible.

It can be shown that, having observed the received signal

$$y(\tau) = \begin{cases} x_0(\tau) + n(\tau) \\ x_1(\tau) + n(\tau) \end{cases} \tag{2}$$

but not knowing which x was added to n, the *a posteriori* probability that, given $y(t)$ a zero was transmitted is

$$P(0\,|\,y(\tau)) = F_1\left(\int_0^{T_D} [x_0(\tau) - y(\tau)]^2\,\mathrm{d}\tau\right), \tag{3}$$

and similarly

$$P(1\,|\,y(\tau)) = F_2\left(\int_0^{T_D} [x_1(\tau) - y(\tau)]^2\,\mathrm{d}\tau\right), \tag{4}$$

whe F_1 and F_2 are monotonic decreasing functions of their arguments. Errors will be as infrequent as it is possible to make them if these two probabilities can be computed and a *decision* made in favor of 0 if $P(0\,|\,y(t)) > P(1\,|\,y(t))$ and in favor of 1 if vice versa. F_1 and F_2 depend on the *a priori* probabilities $P(0)$ and $P(1)$ and if these and the waveforms $x_0(\tau)$ and $x_1(\tau)$ are known at the receiving end of the system, all the data necessary to compute $P(0\,|\,y(t))$ and $P(1\,|\,y(t))$ is at hand, and an optimum receiver can be constructed. The receiver shown in Fig. 2 is such a receiver for the common case in which

$$P(0) = P(1) = \tfrac{1}{2} \tag{5}$$

and the transmitter signal energies are equal

$$\int_0^{T_D} x_0^2(\tau)\,\mathrm{d}\tau = \int_0^{T_D} x_1(\tau)\,\mathrm{d}\tau = E_s\,. \tag{6}$$

Condition (5) means that $F_1 = F_2$ and because this function is monotonic, a decision can be made by comparing just the arguments in equations (3) and (4). By expanding the integrands of these arguments and using (6) it is seen that a decision should be made according to which cross-correlation

$$\int_0^{T_D} x_0(\tau)\,y(\tau)\,\mathrm{d}\tau\,,$$

or

$$\int_0^{T_D} x_1(\tau)\,y(\tau)\,\mathrm{d}\tau\,,$$

is the larger, rather than according to which mean-square difference (equations (3) and (4)) is the smaller.

Having found the optimum receiver we might be tempted to go further and discover some optimum choice of transmitter waveforms $x_0(\tau)$ and $x_1(\tau)$, under an average power constraints $E_s =$ a constant. However, it turns out that an optimum choice for $x_0(\tau)$ and $x_1(\tau)$ is merely that they be equal and opposite. As long as they obey this condition and have energy E_s, the particular choice of waveform does not matter. The probability of error as it turns out, is a function only of E_s/N_0.

Now let us proceed to introduce the multipath condition and see what can be said about the optimum receiver and the best choice of signals to use with it.

The multipath condition, representable by the network of Fig. 1-C, is assumed to be in cascade with the transmitted signal before the noise $n(\tau)$ is added, *i.e.* the filter appears at point P in Fig. 2. The receiver is now no longer optimum since it is correlating $x_0(\tau)$ and $x_1(\tau)$ against

$$y(\tau) = \begin{cases} x_0(\tau) * h(\tau, t) + n(\tau) \\ x_1(\tau) * h(\tau, t) + n(\tau) \end{cases} \tag{7}$$

rather than the $y(\tau)$ given by equation (2). (The $*$ indicates the convolution operation describing the output of a linear filter.) But notice that if the reference signals at the receiver were not $x_0(\tau)$ and $x_1(\tau)$, but rather $x_0(\tau) * h(\tau, t)$ and $x_1(\tau) * h(\tau, t)$ respectively, by the arguments given previously, the receiver would be optimum again. And this can clearly be done by inserting filters, identical to that representing the multipath channel, in cascade with the sources of signals $x_0(\tau)$ and $x_1(\tau)$, as shown in Fig. 3. As the succession of signals lasting T_D are transmitted, and the multipath condition (represented by all the α_i's and Φ_i's) changes slowly, the receiver will still be optimum as long as the α_i's and Φ_i's in the two filters of the receiver are kept in correspondence with those in the channel. (This correspondence is indicated by broken lines in Fig. 3).

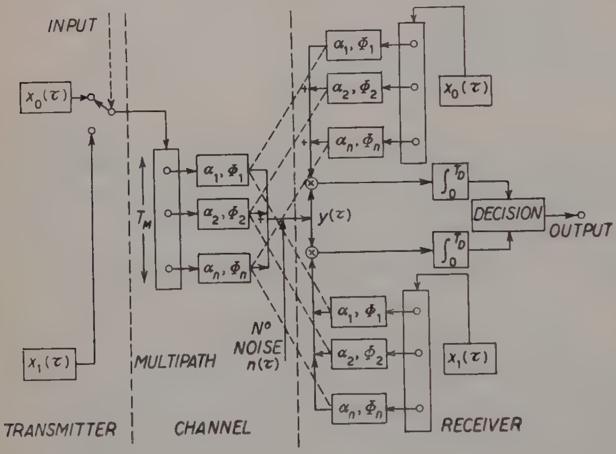


Fig. 3.

Now the question of the optimum transmitter signals is not so easily answerable. The probability of error, as before, depends on the ratio of signal energy E_s' to noise spectral density N_0, where by signal energy we mean the signal at the point where noise is added to it:

$$E_s' = \int_0^{T_D} [x_0(\tau) * h(\tau, t)]^2 \, d\tau \,, \tag{8}$$

and similarly for $x_1(\tau)$. But in a physical system we will want to constrain the transmitter power as indicated in equation (6). To minimize the probability of error under this constraint but with the multipath condition present, it can be easily shown from (8) that $x_0(\tau)$, and $x_1(\tau)$ should be sinewaves of opposite phase at the frequency in W for which $|H(\omega, t)|^2$ is a maximum. For present purposes, however, this solution must be considered to be of academic interest only, since it implies some return link from receiver to transmitter so that the proper wave forms may be mutually agreed upon. Communication systems with such feedback links are interesting but here we will be obliged to confine ourselves arbitrarily to the condition for which the waveforms $x_0(\tau)$ and $x_1(\tau)$, once agreed upon, are not altered. The question of optimum signal waveforms under this condition has not been solved, and as will be seen from subsequent paragraphs the particular choice used by us has been based more or less on intuitive reasoning, and represents, at best, a start on the problem.

Returning to Fig. 3, we perceive that in order for the correspondences indicated by dotted lines to be maintained, the receiver must somehow make measurements of the α_i's and Φ_i's and use them in the correcting filters. Before discussing the measurement function we redraw the receiver of Fig. 3 in the form given in Fig. 4. To do this we note that in Fig. 3 the input to each integrator is obtained by multiplying $y(\tau)$ by the sum of delayed replicas of $x(\tau)$, each replica having been multiplied by an amplitude and phase angle. By doing the multiplication ahead of the weighting and developing variously delayed replicas of $y(\tau)$ rather than of $x(\tau)$ we have the scheme of Fig. 4. That this is legitimate and also has certain practical advantages is shown in the parent paper ([1], p. 561).

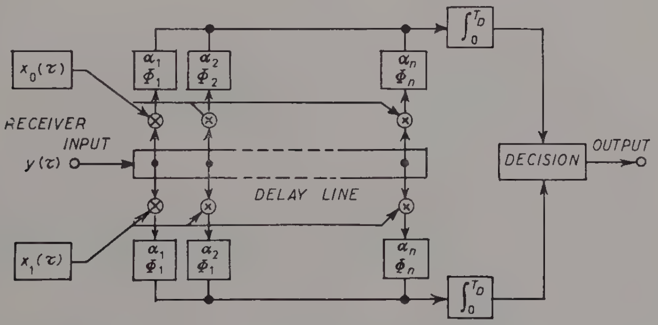


Fig. 4.

The measurement of $h(\tau, t)$, that is the α_i's and Φ_i's, can obviously not be done perfectly, because of the noise. However, remembering that the fluctuation period $1/R$ of these parameters is much longer than T_D (expression (1)) the length of each signalling element, we are tempted to measure the α_i and Φ_i

as accurately as we can by observing for $1/R$ seconds knowing that each measurement will be less noisy than the two inputs to the decision operation, since the latter are allowed an observation interval of only T_D.

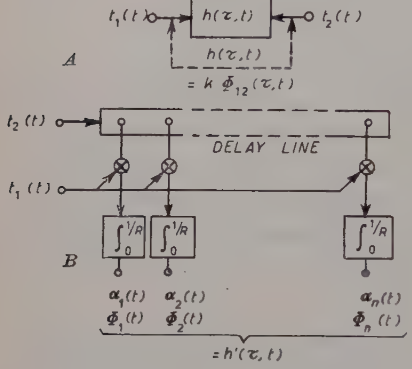


Fig. 5.

There are several different ways in which one can measure the impulsive response of a filter. A particularly appropriate one, shown in Fig. 5-*A*, is the well-known artifice of cross-correlating input and output. For non-time varying filters

$$H(\omega) = \Phi_{12}(\omega)/\Phi_{11}(\omega)\,, \tag{9}$$

where $\Phi_{12}(\omega)$ and $\Phi_{11}(\omega)$ are the Fourier transforms of the input-output cross-correlation function, and input correlation function, respectively. If $\Phi_{11}(\omega)$ is a constant, we have

$$h(\tau) = K\Phi_{12}(\tau)\,. \tag{10}$$

As long as $h(\tau, t)$ and thereby $H(\omega, t)$ are changing slowly enough (see expression (1)) the same method is applicable and we can write approximately

$$h(\tau, t) = K\Phi_{12}(\tau, t)\,. \tag{11}$$

Since we are only interested in $h'(\tau, t)$ the inverse Fourier transform of that part of $H(\omega, t)$ lying in W, we can assume the input to have a non-zero spectrum only in W. Thus we can use the device of Fig. 5-*B* to cross-correlate input and output f_1 and f_2 and deliver sample values α_i and Φ_i which define uniquely the band-limited function $h'(\tau, t)$. Each integrator provides as its output the integral of the past $1/R$ seconds of its input. A careful distinction should be made at this point between Figs. 1-*C* and 5-*B*. The former is an equivalent representation of a filter having a band-limited impulse response. The latter is a device to measure such a response by cross-correlating the output and input of such a filter.

In our communication system the receiver is in possession of the filter output f_2, albeit in a noisy form. All that can be said about the input is that it is either $x_0(\tau)$ or $x_1(\tau)$. We can still make the measurement if we use as f_1 in Fig. 5-*B* the mixture $x_0(\tau)+x_1(\tau)$ and insure that $x_0(\tau)$ and $x_1(\tau)$ are reasonably orthogonal for the integration time $1/R$. Then the measurement outputs α_i and Φ_i will be only slightly more noisy than if the actual input sequence were known.

And now as a final step in deriving the optimum receiver we notice by comparing Figs. 4 and 5-*B* that we can use the delay line and multipliers of the former to get the voltages which when integrated for $1/R$ seconds give the approximate values of the α_i and Φ_i shown in Fig. 5-*B*. This is done in Fig. 6 which shows the output of each upper multiplier being added to that of the lower to form the mixture feeding the $1/R$-second integrator. The output of each such integrator is applied as the α_i and Φ_i correction to the output of the first multiplier.

To recapitulate: we have taken the known result for an optimum receiver in the presence of white gaussian noise and modified it to include the case in which there is also some known multipath disturbance in cascade in the channel. Then we have argued, somewhat intuitively, that if the cascaded filter is changing slowly we can make a reasonably error-free measurement of it at the receiver and apply this knowledge for continuous and automatic readjustment of the filter used in the receiver. The name « rake » derives from the manner in which the various correlation detectors are arranged equidistantly along an axis of delays so as to detect any signal arriving in the range of delays T_M.

Recall that we used as *a priori* information about the multipath condition only that the spread in path delays was smaller than some T_M, that the medium was linear and that in a sampled-type representation of the equivalent linear filter, α_i and Φ_i varied more slowly than some rate $R \ll 1/T_M$: This has allowed a reasonably simple explanation of the rake receiver and its statistically optimum properties.

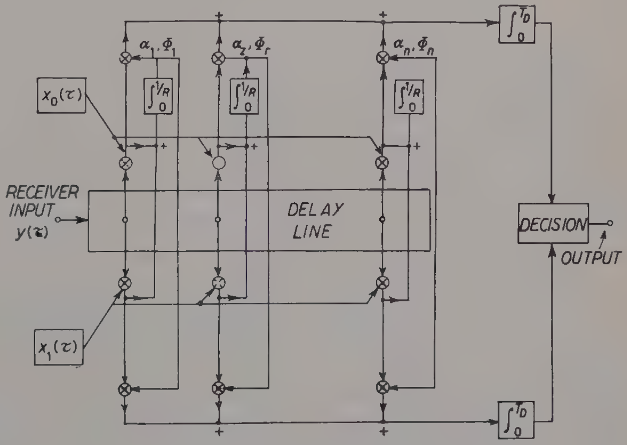


Fig. 6.

This form of receiver was originally derived by PRICE [2] in a different way. He used a considerably more sophisticated *a priori* picture of the propagating medium (specifically the delays, the probability distributions of amplitudes and phases of each of the paths and the power spectrum of their time variation). From this (for large N_0) an optimum receiver somewhat like that in Fig. 6 was derived more or less directly, with the measurement operation already contained in the result, and not introduced *ad hoc* as in the treatment given here. The principal heuristic extension necessary in Price's result was to assume that it still holds for reasonable values of N_0 and to provide taps on the receiver delay line spaced by the sampling interval $(1/W)$, rather than to leave them placed at the mean delay of each of the assumed discrete paths.

The previous paragraph is by way of historical perspective and points out that we have been unable to construct a unified derivation for the ıake receiver that has not required some supplemental reasoning of a heuristic nature.

The question of optimum wave shapes for $x_0(\tau)$ and $x_1(\tau)$ is worth a few more remarks. It has been noted that the two should be roughly orthogonal for the integration time $1/R$ in order to permit the measurement function and it will be recognized that they should also be orthogonal over the integration time T_D as well, so as to minimize the probability of error. We have referred repeatedly to a bandwidth W without stating what it should be, and we have stated that the spectrum should be flat within W. Using known results on the signal-to-noise ratio observed at the output of correlation detectors ([3], equation (13)) it is possible to derive the following approximate expression for signal to-noise ratio at the decision element input of the receiver of Fig. 6 as a function of the parameters α_i defining the multipath condition

$$\left(\frac{S}{N}\right) = T_D W \frac{\sum_{i=1}^{T_M W} a_i^2}{N_0 W + \sum_{i=1}^{T_M W} a_i^2} . \tag{12}$$

This expression assumes that the signals are segments of gaussian noise of flat density in W. The first denominator term shows the effect of the additive channel noise $n(\tau)$ whereas the second is a *self-noise* term which can be reduced by choosing the waveform statistics to be something other than gaussian. It is not known exactly how large a reduction in this term is theoretically achievable, nor what waveforms to use in achieving it.

From equation (12) it is seen that the proper choice of W is to make it as large as possible. When the first denominator term is smaller than the second (large receiver input signal-to-noise ratio), (S/N) is proportional to the coefficient $T_D W$. When the first term is greater than the second (small input signal-to-noise ratio) an increase in W adds more terms to the numerator summation. This continues until all paths have been resolved whereupon a further increase of W no longer helps. There is experimental evidence that in the high frequency band this condition is not achieved for bandwidths less than 50 to 100 kHz, and one does not usually have this much bandwidth at his disposal.

A system employing the ıake receiver has been built and subjected to limited field tests, using $W = 10$ kHz, $T_M = 3$ ms, $T_D = 22$ ms and R (in Fig. 6)

$=1$ Hz The results of these tests and comparisons with conventional systems are described in reference [1]. The system compares favorably with the more conventional FSK (frequency-shift keying) systems using space-diversity particularly at low error rates. This is what would be expected since the wide-band signal used in the ıake system, and its optimum reception constitutes a form of optimum frequency diversity. Its most important practical limitation (besides equipment complexity) is its use of a wider bandwidth than is usually available. However, situations can be imagined in which the extremely low error rates achievable with the wide-band rake technique would be worth the expenditure of bandwidth.

Perhaps the most serious remaining problem in this work is the one that I have alluded to repeatedly — the study of more suitable waveforms. The size of the second denominator of equation (12) has been observed to be a limiting factor in system performance. Until this problem is solved the full potentialities of the system will not be realized.

REFERENCES

[1] R. Price and P. E. Green jr.: *IRE Proc.*, **46**, 555 (1958).
[2] R. Price: *IRE Trans.*, Vol. IT-2, 125 (1956).
[3] P. E. Green jr.: *IRE Trans.*, Vol. IT-3, 10 (1957).

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3° Trimestre

Network Theoretical and Physical Limitations of Amplifier Noise Performance.

H. A. Haus

Massachusetts Institute of Technology - Cambridge, Mass.

1. – Introduction.

An information system is often represented schematically as a cascade of an information source, an encoder, a channel perturbed by noise, a decoder and an information sink (Fig. 1). The noise is usually considered to be introduced in the channel, whereas the processes of encoding and decoding are taken to be free of noise. Such an assumption is often equivalent to the supposition that the actual physical transmission represented by the channel, is followed by an amplification with a high gain so that any additional noise introduced in the physical decoder is negligible compared to the signal level in the decoder. Since such amplifiers introduce noise of their own, which may be comparable to the signal level, they must be considered to be part of the channel in the schematic of the information system of Fig. 1.

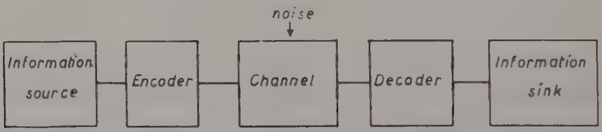


Fig. 1. – Schematic of information system.

We shall be concerned with the optimization of the two terminal-pair amplifiers that have to be employed in order to raise the power level of the signal before its entry into the decoder. In optimizing the operation of these amplifiers we shall make use of the measure of noise performance commonly employed in engineering practice, the noise figure. The noise figure is defined as the quotient of the signal-to-noise ratio at the input of the amplifier to that at the output of the amplifier. Both signal power and noise power are assumed to be contained within a band of frequencies so narrow that the amplifier characteristics may be considered to be constant over the band. In other words the signal to noise ratios are obtained as ratios of spectral densities. The noise-input is assumed to be thermal noise corresponding to standard noise temperature, $T_0 = 290$ °K. The use of the noise figure to characterize amplifier noise performance restricts the problem to a study of single frequency noise performance.

An alternate way of expressing the noise figure F is

$$F = \frac{N}{N_0}, \tag{1}$$

where N is the available output power of the amplifier within a narrow frequency interval Δf, N_0 is the available output power that would exist if the amplifier were noise free.

In the noise figure definition it is implied that the internal noise of the amplifier is additive to the signal passing through the amplifier. We shall assume that this condition of linearity is satisfied throughout the analysis.

We are here concerned with the study of the basic limits of two terminal-pair amplifier noise performance as measured by the noise figure at high gain. The restriction to high gain is a necessary one, since otherwise the problem is not defined. Indeed, the noise figure of any amplifier can be reduced to unity at a complete sacrifice of gain by short-circuiting the input terminals to the output terminals.

With the recognition that amplifiers, basically, provide « gain building blocks » of which it is desired that they add as little as possible to the system noise we shall use the following criterion for the evaluation of quality of amplifier noise performance:

Suppose that n different types of amplifiers are compared. An unlimited number of amplifiers of each type is presumed to be available. A general *lossless* (possibly non-reciprocal) interconnection of an arbitrary number of amplifiers of each type is then envisioned, with terminals so arranged that in each case an over-all *two terminal-pair network* is achieved. For each amplifier type, both the lossless interconnecting network and the number of amplifiers are varied in all possible ways to produce two conditions simultaneously:

a) a very high available gain (approaching infinity) for the over-all two terminal pair system when driven from a source having a positive internal impedance; and

b) an absolute minimum noise figure $F_{\min}$ for the resulting high-gain system.

The value of ($F_{\min}-1$) *for the resulting high-gain two terminal-pair network is taken specifically as the « measure of quality » of the amplifier type in each case. The « best » amplifier type will be the one yielding the smallest value of* ($F_{\min}-1$) *at very high gain.*

The network interconnection envisaged with each amplifier type is shown in Fig. 2. Lossless interconnections are used since only such interconnections do not add to system noise and therefore such interconnections should be able to achieve optimum noise performance. (A proof has been presented [1] to the effect that interconnections with loss cannot lead to a noise performance better than that achievable with lossless interconnections).

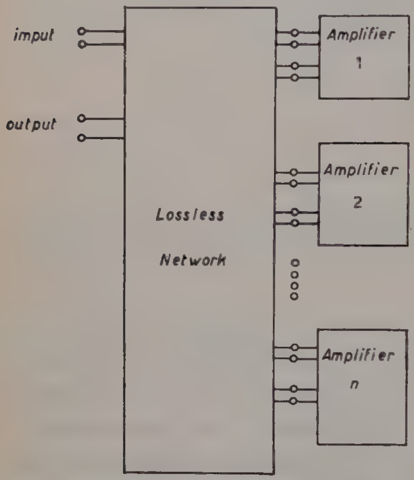


Fig. 2. – Lossless interconnection of amplifiers.

Most of the detailed proofs mentioned here are published elsewhere [1-3]. Here we shall concentrate on three major ideas, which help towards an understanding of the limits on amplifier noisepe rformance.

1) Every two terminal-pair linear noisy network possesses at most two invariants with regard to lossless network transformations performed on the network which leave the number of terminal pairs of the network unchanged. The most general such transformation is shown in Fig. 3. These two invariants are characteristic of the internal noise of the network and of the ability of the network to deliver or absorb power. They represent a convenient summary of all the properties of the network that remain unaffected by lossless transformations.

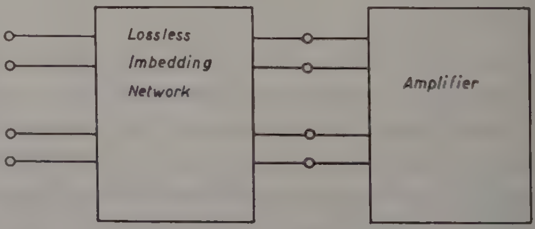


Fig. 3. – Lossless transformation or « imbedding » of two terminal-pair amplifier.

2) The two basic invariants are related to the optimum noise performance achievable with a two terminal-pair amplifier. This relation establishes two facts: *a*) every amplifier possesses a basic limit of its noise performance and *b*) the limit of this noise performance has invariant characteristics (as one would expect from a basic quantity).

3) In some physical cases studied so far there is a connection between the network-theoretical limitation of the noise performance of an amplifier on one hand, and the physical gain mechanisms and noise processes on the other hand. The newly developed maser amplifier will serve as an illustration.

2. – Canonical form of linear noisy two terminal-pair network.

A linear two terminal-pair network with internal sources is conveniently characterized in terms of the impedance matrix relation [*]

$$\underset{\sim}{V}+\underset{\sim}{Z}\,\underset{\sim}{I}=\underset{\sim}{E} \tag{2}$$

Here $\underset{\sim}{V}$ is a column matrix of 2nd order comprising the two terminal voltages of the network, $\underset{\sim}{I}$ comprises the two currents, $\underset{\sim}{Z}$ is a square matrix of 2nd order, and $\underset{\sim}{E}$ comprises the two open-circuit noise voltages of the network. In the study of noise the complex amplitudes of the voltages are meaningless by themselves and only self and cross-spectral densities have physical significance. They are conveniently summarized in a square matrix of 2nd order which is

$$\overline{\underset{\sim}{E}\,\underset{\sim}{E}^{\dagger}}=\begin{bmatrix}\overline{E_1E_1^*} & \overline{E_1E_2^*}\\ \overline{E_1^*E_2} & \overline{E_2E_2^*}\end{bmatrix}. \tag{3}$$

Here the dagger † indicates the operation of taking the complex conjugate transpose of a matrix. The complex amplitudes $\underset{\sim}{E}$ are supposed to be RMS. A linear noisy two terminal-pair network is completely characterized at any particular frequency by the two matrices $\underset{\sim}{Z}$ and $\overline{\underset{\sim}{E}\underset{\sim}{E}^{\dagger}}$.

If a lossless network transformation is performed on the network such as shown in Fig. 3, henceforth called an « imbedding », a new network is obtained with a new impedance matrix and a new correlation matrix $\overline{\underset{\sim}{E}\underset{\sim}{E}^{\dagger}}$. This shows that the eight parameters (six of which are complex) characterizing a particular two terminal-pair network may be affected by a lossless transformation. One may suspect that at least some features of these eight parameters ought to be preserved in such a transformation. In other words, one would expect that every network possesses a certain set of invariants with regard to lossless transformations.

In references [1] and [2], R. B. ADLER and the author indeed found that every two terminal-pair network possesses all in all two invariants (one of

[*] We concentrate on two terminal-pair networks since these are used as amplifiers. The proofs have actually been carried out for n terminal pair networks.

which may assume the trivial value of zero). These two invariants are the eigenvalues of the 2nd order characteristic noise matrix defined by

$$\underset{\sim}{N} = -\tfrac{1}{2}(\underset{\sim}{Z} + \underset{\sim}{Z}^{\dagger})^{-1}\overline{\underset{\sim}{E}\underset{\sim}{E}^{\dagger}}\,. \tag{4}$$

The characteristic noise matrix (4) contains two significant features of a linear noisy network. First, there is the positive definite correlation matrix $\overline{\underset{\sim}{E}\underset{\sim}{E}^{\dagger}}$ which describes the noise within the network. Secondly, it contains the inverse of the matrix $(\underset{\sim}{Z}+\underset{\sim}{Z}^{\dagger})$ which characterizes the ability of the network to generate or absorb power. Indeed, in the absence of noise, the power P entering the network is given by

$$P = \tfrac{1}{2}\underset{\sim}{I}^{\dagger}(\underset{\sim}{Z} + \underset{\sim}{Z}^{\dagger})\underset{\sim}{I}\,. \tag{5}$$

In the classification of networks three cases have to be distinguished:

a) The network is passive, $\underset{\sim}{Z}+\underset{\sim}{Z}^{\dagger}$ is positive definite.

b) The network is incapable of power absorption and generates power under any arbitrary adjustment of the terminal currents. The matrix $\underset{\sim}{Z}+\underset{\sim}{Z}^{\dagger}$ is negative-definite. This is the case of a negative resistance network.

c) The matrix $\underset{\sim}{Z}+\underset{\sim}{Z}^{\dagger}$ is indefinite. The network can either generate or absorb power depending upon the adjustment of the terminal currents.

The signs of the eigenvalues of $\underset{\sim}{N}$ in Eq. (4) can now be determined from the fact that the signature of $\underset{\sim}{N}$ is controlled by the signature of $\underset{\sim}{Z}+\underset{\sim}{Z}^{\dagger}$, since the correlation matrix is positive definite. For the three cases distinguished above, we have

a) both eigenvalues are negative,

b) both eigenvalues are positive,

c) the two eigenvalues are of opposite sign.

The fact that every two terminal-pair network has 2 invariants with respect to lossless transformations suggests that there should existat least one lossless transformation for every particular two terminal-pair network that reduces the network into a form which places the two invariants into direct evidence. This « canonical » form should not contain more than two parameters representing the two invariants of the network. A proof to that effect

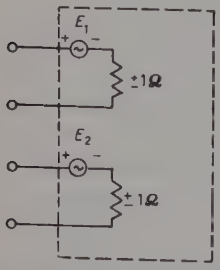


Fig. 4. – Canonical form of two terminal-pair network. The signs of the resistances are for case (*a*) + + ; (*b*) — — ; (*c*) + —.

has been carried out [2] and the resulting canonical form of the network is represented in Fig. 4.

The proof is summarized in the following theorem: At any particular frequency, every two terminal-pair network can be reduced by lossless imbedding into a canonical form consisting of two separate (possibly negative) resistances in series with uncorrelated noise voltage generators $\overline{|E_i|^2}$.

The two values of the $\overline{|E_i|^2}$ are related to the two eigenvalues λ_i of the characteristic noise matrix $\underset{\sim}{N}$ by the formula

$$\overline{|E_i|^2} = \pm \lambda_i , \qquad i = 1, 2. \tag{6}$$

where the — sign applies to noise voltages pertaining to a positive resistance, the + sign to those pertaining to a negative resistance. The sign of the two resistances that appear in the canonical form of a network are uniquely determined by the impedance matrix of the original network and *a*) are both positive for a passive network, *b*) are both negative for a negative resistance network, *c*) one is positive, and the other is negative, for a network with an indefinite $\underset{\sim}{Z}+\underset{\sim}{Z}^\dagger$ matrix.

Equation (6) can be checked easily by evaluation of the characteristic noise matrix for the canonical network. In this case $\underset{\sim}{N}$ is diagonal.

In connection with the above theorem it is worth noting that the eigenvalues of $\underset{\sim}{N}$ for a passive network at thermal equilibrium with the equilibrium temperature T are all identical and have the value $-kT\Delta f$. This statement is rather obvious in connection with Eq. (6) and the fact that each positive resistance of the canonical form must have an available power of $kT\Delta f$ according to the Nyquist formula.

A canonical form is particularly convenient in summarizing the essential unalterable characteristics of a linear noisy network. The canonical form of the amplifier is also helpful to recognize the network theoretical limitations on the noise performance of a linear amplifier.

3. – Amplification as a coupling to a negative resistance.

The fact that every two terminal-pair network possesses a canonical form as shown in Fig. 4 may now be used to obtain an understanding of the reasons for the existence of a basic limit on amplifier noise performance. Thus, consider the problem of noise figure optimization as originally stated and illustrated in Fig. 2. The imbedding network is assumed to be the general network that leads to optimum noise performance. Every amplifier can be represented by its canonical form imbedded in the lossless network which is inverse to that needed to reduce the original amplifier into canonical form. The lossless imbedding networks may be considered to be part of the general imbedding network of Fig. 2. One then obtains for Fig. 2 the new Fig. 5. For definiteness we have assumed that our amplifiers are all identical, with $\underset{\sim}{Z}+\underset{\sim}{Z}^{\dagger}$ indefinite. We shall adhere to this assumption throughout and only state at the end of the discussion how our arguments have to be modified in order to take into account the class of amplifiers with $\underset{\sim}{Z}+\underset{\sim}{Z}^{\dagger}$ negative definite.

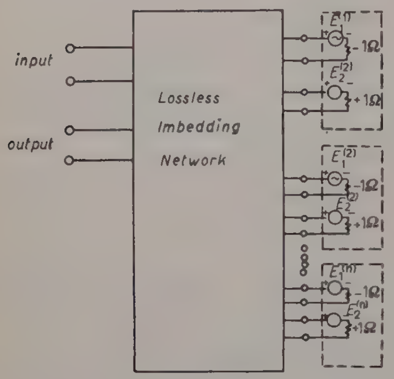


Fig. 5. – Alternative representation of lossless interconnection of amplifiers.

Next, we shall demonstrate that the most common conventional amplifier, with the equivalent circuit shown in Fig. 6, may be considered to be composed of a negative resistance, a positive resistance, and a lossless nonreciprocal device, a circulator.

The negative resistance provides the gain, the positive resistance accounts for the dissipative behavior of the network of Fig. 6 when excited from terminal-pair 3. The circulator is a lossless non-reciprocal network that transmits waves incident on any one of its four terminal-pairs as shown in Fig. 7. The transmission lines connected to the four terminal-pairs are assumed to have one ohm characteristic impedance. If a wave is incident on transmission line (1) of the circulator and if all other three terminal-pairs are matched with resistances of one ohm, the wave is transmitted without loss to terminals (2). On the other hand, if a wave is incident from terminals (2) with the remaining three terminal-pairs matched, the wave is transmitted without loss to terminals (3), and so forth.

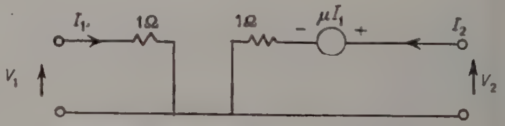


Fig. 6. – Unilateral amplifier.

We connect a negative resistance of $-1\ \Omega$ through an ideal transformer of turns-ratio $m:1$ to terminals (2) of Fig. 7. The resistance seen on the secondary of the transformer is then

$$R=-m^2\,.$$

On terminals (4) we connect a positive resistance of $+1\ \Omega$. Terminals (1) are used as the input of the amplifier, terminals (3) as the output. We shall now inspect the nature of the equivalent circuit of the resulting amplifier disregarding the internal noise for the time being.

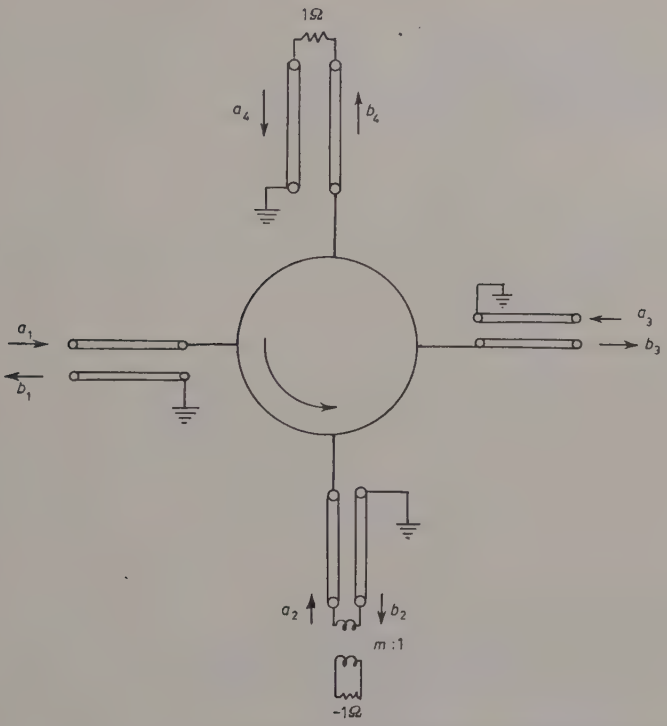


Fig. 7. – Circulator with positive and negative resistance connected

Suppose a wave is incident from the input on terminals (1). This wave is transmitted to the negative resistance connected to terminals (2) and reflected with a corresponding increase of power.

$$a_2 = \Gamma b \;= \Gamma a_1\,, \tag{7}$$

where

$$\Gamma = \frac{R-1}{R+1}\,. \tag{8}$$

Γ is the reflection coefficient of the resistance R, a quantity of magnitude greater than unity since R is negative. The wave a_2 appears without loss at terminals 3. We thus have

$$b_3 = \Gamma a_1\,; \qquad \text{for} \quad a_3 = 0 \tag{9}$$

and since no reflected wave appears at terminals 1 we have

$$b_1 = 0 \; ; \qquad \text{for} \quad a_3 = 0 \; . \tag{10}$$

Now applying another boundary condition to the device by setting $a_3 \neq 0$ and $a_1 = 0$, we find that the wave incident upon terminals (3) is transmitted directly into terminals (4). We thus have

$$b_1 = 0 \qquad \text{for} \qquad a_1 = 0 \; . \tag{11}$$

Comprising Eqs. (9) to (11) into a single scattering matrix relationship we find

$$b_3 = \Gamma a_1$$

$$b_1 = 0$$

and thus see that the network has the scattering matrix

$$\underset{\sim}{S} = \begin{bmatrix} 0 & 0 \\ \Gamma & 0 \end{bmatrix} . \tag{12}$$

The power gain of the amplifier is

$$G = \left| \frac{b_3}{a_1} \right|^2 = |\Gamma|^2 = \left| \frac{R-1}{R+1} \right|^2 . \tag{13}$$

The equivalent circuit of the network with the scattering matrix (12) is that shown in Fig. 6 and is identical with the equivalent circuit of a conventional unilateral amplifier as represented by an ideal triode with finite grid resistance. For μ of Fig. 6 we have

$$\mu = 2 \frac{R-1}{R+1} . \tag{14}$$

The construction employed here demonstrates the nature of *amplification* in a conventional unilateral amplifier with the equivalent circuit of Fig. 6. Amplification is obtained by coupling the input excitation into a negative resistance. The positive resistance connected to the ideal circulator only serves to isolate the input and the output of the amplifier by absorbing any wave incident into the output terminals of the amplifier.

This picture of the gain mechanism is very useful in gaining an understanding of the basic limit on noise performance as proved in references [1-3]. Indeed, returning to the equivalent circuit of Fig. 5 we note that the obtaining of gain from the resulting two terminal-pair network depends upon our ability to couple to at least one of the negative resistances on the right hand side of the circuit. In coupling to this resistance one has to couple to the internal noise of the resistance as well. Since all negative resistances and their noise sources are identical, one may couple to any number as well as to a single one of the entire set. If it is desired to obtain an amplifier with the equivalent circuit of Fig. 6 it is also necessary to couple to at least one of the positive resistances in order to obtain absorption of the power reflected back in the output of the amplifier.

This reasoning leads one to suspect that the network of Fig. 7 is one of the physical forms of the imbedding network of Fig. 5 which realizes the optimum noise performance for the amplifier interconnection. In this form, terminals marked « input » and « output » in Fig. 5 are taken to correspond to terminals (1) and (3) of the circulator. Terminals (2) and (4) of the circulator are any other two of the terminal-pairs of the « imbedding network », one connected to positive resistance the other to a negative resistance. That this particular network form indeed realizes the optimum noise performance will be confirmed by direct evaluation. For this purpose, it is helpful to introduce a new measure of noise performance which preserves its significance even at low amplifier gain whereas the noise figure serves as a noise performance criterion only at high gains.

4. – A quantitative measure of amplifier noise performance for amplifiers of low gain.

We have accepted as the measure of quality of noise performance the minimum noise figure at high gain that can be achieved by a lossless interconnection of amplifiers of a given type. In the detailed study reported elsewhere it was found helpful to introduce an auxiliary measure of noise performance.

The measure of noise performance is

$$M = \frac{F-1}{1-(1/G)}\,. \tag{14a}$$

Here, G is the available gain of the two terminal-pair network (for a network with loss G is less than unity). Certain modifications in the above definition are necessary when the source impedance used, or the output impedance of the amplifier, have negative real parts. Here we shall not be concerned with such cases.

It is clear from the definition of M that for high gain ($G \to \infty$) it reduces to the excess noise figure. In this limiting case, M serves directly as the measure of noise performance previously accepted. At small gains it can be shown that M has a significance of its own so that it can also be accepted as an appropriate measure of noise performance. M has the following interesting properties:

1) In a cascade of two amplifiers with different noise measures the amplifier with the least noise measure (and not necessarily noise figure) should be used as the first stage in order to obtain the least overall noise figure.

2) A cascade of amplifiers with identical noise measures (but not necessarily the same gain and noise figures) leads to an amplifier with the same noise measure.

3) The smallest value of the noise measure achievable by lossless or passive transformations performed on this amplifier is given by

$$M_{\text{opt}} = \frac{\lambda_1}{kT\,\Delta f}\,, \tag{15}$$

where λ_1, is the least positive eigenvalue of the characteristic noise matrix (4).

Thus M has all the properties one would require from a fundamental measure of amplifier noise performance.

5. – Realization of the optimum noise performance.

We shall now illustrate the realization of the lower limit on noise performance with the aid of the circulator scheme discussed in Sect. **3**. We connect the negative resistance of the canonical form of the amplifier belonging to class c) to terminals (2) of the circulator through an ideal transformer of turns ratio $m:1$, the positive resistance to terminals (4). We have for the noise figure

$$F = \frac{\overline{|b_3|^2}}{\overline{|\Gamma a_1|^2}}\,, \tag{16}$$

where one takes $\overline{|a_1|^2} = kT\,\Delta f$.

In b_3 of Eq. (16) appears the noise internal to the amplifier. We have

$$\overline{b_3^2} = \overline{\left|\Gamma a_1 + \frac{E'}{1+R}\right|^2}\,, \tag{17}$$

where E' is the voltage seen from the secondary of the ideal transformer, $E' = mE_1$.

The noise measure (14) thus becomes

$$M = \frac{\overline{|E'|^2}}{|\Gamma|^2(1+R)^2\,kT\,\Delta f[1-((1+R)/1-R)^2]} = \tag{18}$$

$$= -\frac{\overline{|E'|^2}}{4RkT\,\Delta f} = \frac{\overline{|E_1|^2}}{4kT\,\Delta f} = \frac{\lambda_1}{4kT\,\Delta f}\,,$$

where λ_1 is the positive eigenvalue of the characteristic noise matrix of the amplifier. Thus, the scheme of Fig. 7 realizes the lowest possible value of M at *every adjustment of the ideal transformer*, in particular for $m \to 1$, *i.e.* for the limit of $G \to \infty$ (see Eq. (14)).

In the optimization of noise performance carried out so far it was assumed that the amplifiers belonged to class (*c*), so that the canonical form contained both a positive and a negative resistance. The negative resistance was connected to terminal-pair (2) of the circulator and provided the gain, and the positive resistance was connected to terminal-pair (4) in order to provide the absorption of a wave incident into the output terminal pair of the amplifier. We shall now discuss the modifications necessary in the optimization scheme when using amplifiers of class (*b*) with two negative resistances in their canonical form. Clearly, when optimizing the noise performance, power gain should be obtained by coupling to that of the two resistances which has the smaller open-circuit noise voltage. In the scheme employing a circulator in connection with one positive and one negative resistance, one uses this negative resistance on terminal-pair (2). On terminal-pair (4) one may connect any positive resistance. The noise of this resistance does not affect the noise figure since the noise is all absorbed in the source. The optimum noise measure is again given by

$$M_{\text{opt}} = \frac{|\overline{E_1^2}|}{4kT\,\Delta f},$$

where $|\overline{E_1^2}|$ is the open circuit noise voltage of the negative resistance used in the amplifier. This particular noise-measure optimization scheme employs (essentially) a lossy network and is, therefore, an example of a noise performance optimization with a lossy network.

It is interesting to note that the noise mesure of a passive network, at the equilibrium temperature T_0 of its source, is always -1. Indeed, the power available at the network output terminals must satisfy the Nyquist formula, but is also directly related to the noise figure by definition. One has

$$N = FGkT_0\Delta f = kT_0\Delta f$$

or

$$F = 1/G$$

thus

$$M = \frac{F-1}{1-(1/G)} = -1\,. \tag{19}$$

For a network at temperature T one has instead

$$M = -\frac{T}{T_0}\,. \tag{20}$$

6. – Physical limitations of amplifier noise performance.

We shall now turn to a brief study how the lower limits to amplifier noise performance are established physically. The limits are understood in several cases, such as the general microwave electron beam amplifier [4], of which the conventional triode is a special case, the new parametric amplifiers and the maser amplifier. Although the triode is the most common amplifier, we shall not study its noise performance here since its study is rather involved. Instead, we shall study a somewhat over-simplified version of a two-level maser in which the limiting noise performance and thermodynamics are particularly intimately connected.

In a paramagnetic salt with two quantum mechanical energy levels, the ratio of the populations of states in the upper and the lower levels at equilibrium temperature T is given by the well known Boltzmann formula

$$\frac{n_u}{n_l} = \exp\left[-\frac{h\nu}{kT}\right] \tag{21}$$

where $h\nu$ is the separation between the energy levels measured in terms of frequency ν, and T is the temperature of the salt. This situation is illustrated schematically in Fig. 8 where the exponential factor is sketched.

A salt with the population distribution of Fig. 8 is passive and has net absorption of radiation incident upon it. Suppose now that the populations in the two energy levels are reversed. This can be accomplished in principle in a paramagnetic salt by a fast reversal of the time average magnetic field, if the energy separation is originally caused by such a field. The spin systems then do not change their orientation fast enough but remain, for some time, in their original spatial orientation. The distribution of states is now

$$\frac{n_u}{n_l} = \exp\left[\frac{h\nu}{kT}\right]. \tag{22}$$

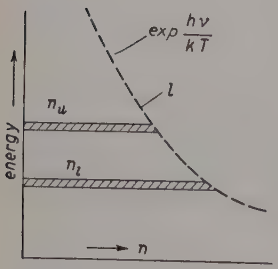


Fig. 8. – Populations in the two energy levels.

The material is now emissive and « presents a negative resistance » to radiation. It is noteworthy that the original temperature of the sample appears in the ratio (22) with a reversal of sign in the exponent.

If the form of the Boltzmann factor is retained even for the active state of the salt, one may characterize the material by a negative temperature $T_e = -T$.

While such a practice is debatable, it is still interesting that this same negative temperature plays an important role in determining the basic limit on the noise performance of the sample. One can show quite in general that the best noise measure achievable with the two-level maser just described is given by

$$M_{\text{opt}} = -\frac{T_e}{T_0},$$

where T_e is the *negative* temperature appearing in the Boltzman factor characterizing the distribution of states of the excited two-level maser. This expression should be compared with the expression for the noise measure of a passive network, Eq. (20).

In summary, we may state that the answer for the basic limit on the noise performance of a two-level maser was particularly simple since we had to deal with a state of inverted thermal equilibrium which still retains many characteristics of thermal equilibrium. For thermal equilibrium, however, the noise measure defined here acquires a particularly simple value.

REFERENCES

[1] H. A. Haus and R. B. Adler: *Optimum Noise Performance of Linear Amplifiers*, in *IRE Proc.*, **46**, 1517 (1958).

[2] H. A. Haus and R. B. Adler: *Canonical Form of Linear Noisy Networks*, in *IRE Trans.*, Vol. CT-5, 161 (September 1958).

[3] R. B. Adler and H. A. Haus: *Network Realization of Optimum Noise Performance*, in*IRE Trans.*, Vol. CT-5, 156 (September 1958).

[4] H. A. Haus and F. N. H. Robinson: *The minimum Noise Figure of Microwave Beam Amplifiers*, in *IRE Proc.*, **43**, 981 (August 1955).

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3° Trimestre

Statistical Filtering and Prediction.

Y. W. Lee

*Massachusetts Institute of Technology, Research Laboratory of Electronics
Cambridge, Mass.*

1. – Introduction.

The three lectures I shall give in this Course on Information Theory concern the theory of filtering and prediction that was originated by N. Wiener and published by him in the book *Extrapolation, Interpolation and Smoothing of Stationary Time Series* (New York, 1950).

In this theory the messages and noise are assumed to be continuous stationary random processes for which autocorrelation functions exist. If $f_a(t)$ is a message or a noise, where t represents time, then its autocorrelation function is defined as

$$\varphi_{aa}(\tau) = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} f_a(t)\, f_a(t+\tau)\,\mathrm{d}t\,, \tag{1}$$

in which the time displacement τ has the range $(-\infty,\ \infty)$.

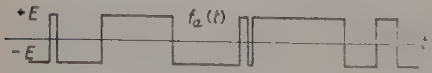


Fig. 1.

For an example of this function consider the rectangular wave $f_a(t)$ shown in Fig. 1. This wave has two possible values of amplitude, namely, $+E$ and $-E$.

We assume that the zero-crossings follow the Poisson distribution

$$P(n,\tau) = \frac{(k\tau)^n}{n!}\exp[-k\tau]\,, \tag{2}$$

which gives the probability of finding n zero-crossings in the duration τ in terms of the average number of zero crossings per second k. The computation of an autocorrelation function is a fairly long story which we shall not be able to tell here. We merely state that the autocorrelation function of the Poisson rectangular wave can be shown to have the expression

$$\varphi_{aa}(\tau) = E^2 \exp[-2k|\tau|] . \tag{3}$$

In a manner similar to (1) we define a crosscorrelation function. Thus if $f_a(t)$ and $f_b(t)$ are two stationary random processes, their crosscorrelation function is defined as

$$\varphi_{ab}(\tau) = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} f_a(t) f_b(t+\tau)\, dt . \tag{4}$$

It is important to note that the second subscript in $\varphi_{ab}(\tau)$ corresponds to the random process that has been given the positive displacement τ. We can show that

$$\varphi_{ab}(\tau) = \varphi_{ba}(-\tau) . \tag{5}$$

Since we shall consider filtering and prediction by means of a linear system, it is necessarily important at this point to state the relationship among the input, output and the characterizing function of the system. It is well known that the relationship in question is the convolution integral

$$f_0(t) = \int_{-\infty}^{\infty} h(\tau)\, f_i(t-\tau)\, d\tau , \tag{6}$$

where $f_i(t)$ and $f_0(t)$ are the input and output, respectively, of the linear system, and $h(t)$ is the time response of the system to a unit-impulse excitation.

2. – Formulation of the problem.

We shall formulate the problem in fairly general terms so that filtering and prediction are particular applications of the general theory. Let us consider Fig. 2 where A represents a linear system which is characterized by $h(t)$, and $f_i(t)$ is its input.

Choosing a particular situation, but not restricting the theory to it, we consider the case where the input is the sum of a message $f_m(t)$ and a noise $f_n(t)$. Thus $f_i(t) = f_m(t) + f_n(t)$. We draw portions of the random processes as shown. To indicate the present time, a peak in the message has been chosen for convenience.

Now, if filtering is our objective, we will state that the desired output of the system A, in the ideal situation, although it cannot be achieved perfectly, is the original message without the noise as shown in the upper-half of the right-hand side of the figure. Frequently, in filtering, a lag in the output message is not a distortion so that in specifying the desired output we write

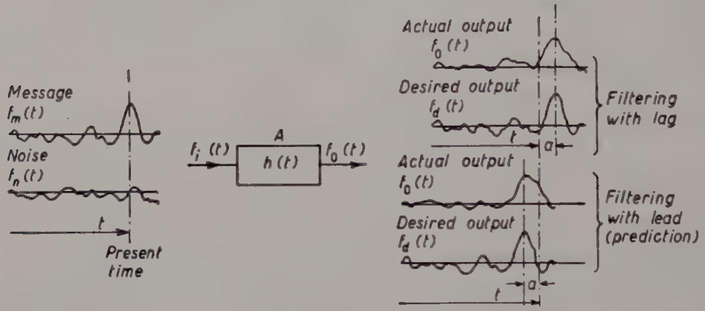


Fig. 2.

$$f_d(t) = f_m(t - \alpha) \, , \tag{7}$$

where $\alpha \geqslant 0$ is the lag. However, in the presence of the noise at the input, the actual out cannot be without error no matter how we may design the linear system. Thus we indicate together with the desired output the actual output $f_0(t)$ which we shall attempt to make as close as possible to $f_d(t)$, based upon a chosen criterion, by properly designing the system.

When a lag in the output message is undesirable, but a lead in it is an advantage, as in control problems, we then specify that the desired output is the original input message with a forward displacement in time. This specification combines filtering with prediction as we can readily see. The desired output is therefore

$$f_d(t) = f_m(t + \alpha) \, , \tag{8}$$

where $\alpha \geqslant 0$ is the prediction time. We have indicated this in the figure where the past is to the left of the present, and the future is to the right of it. The actual output $f_0(t)$ is shown in a manner similar to the case of filtering with lag.

As another example, consider pure prediction. By pure prediction we mean the prediction of a message in the absence of a disturbance. For this problem we have

$$f_i(t) = f_m(t) \tag{9}$$

and

$$f_d(t) = f_m(t+\alpha)\,. \tag{10}$$

We have give only a few examples for illustrating the idea of a desired output in a problem. This idea combines operations such as filtering and prediction, which are conceived to be entirely unrelated in conventional theory, into one single problem. At a later stage we shall discuss a further generalization of the idea.

Concerning prediction we should remember that the correlation functions upon which this theory is based are functions derived from statistical descriptions of the messages and noise involevd in the problem. The prediction is therefore a statistical prediction by means of a linear operation.

Having specified the input and desired output of a linear system we are ready to consider the performance of the linear system and the method of finding that system which gives the best performance. The instantaneous error is clearly the difference between the actual output and the desired output. It is desirable that the measure of error is always positive for any instantaneous error, and is mathematically manageable. Such a measure is the mean-square error defined as

$$\varepsilon = \lim_{T\to\infty}\frac{1}{2T}\int\limits_{-T}^{T}[f_0(t)-f_d(t)]^2\,\mathrm{d}t\,, \tag{11}$$

which is simply the mean square of the instantaneous difference between the actual output and the desired output.

We are now interested in the relationship between the mean-square error and the correlation functions which characterize the input and the desired output. To find the relationship, we write the convolution integral (6) for $f_0(t)$ in (11) so that the mean square error is brought into relation with the system characteristic and the input. Thus

$$\varepsilon = \lim_{T\to\infty}\frac{1}{2T}\int\limits_{-T}^{T}\left[\int\limits_{-\infty}^{\infty}h(\tau)f_i(t-\tau)\,\mathrm{d}\tau - f_d(t)\right]^2\mathrm{d}t\,. \tag{12}$$

Expanding the expression we have

$$(13)\qquad \varepsilon = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}\mathrm{d}t\left[\int_{-\infty}^{\infty}h(\tau)\,f_i(t-\tau)\,\mathrm{d}\tau\int_{-\infty}^{\infty}h(\sigma)\,f_i(t-\sigma)\,\mathrm{d}\sigma - {} - 2f_d(t)\int_{-\infty}^{\infty}h(\tau)\,f_i(t-\tau)\,\mathrm{d}\tau + f_d^2(t)\right].$$

Inverting orders of integration we have

$$(14)\qquad \varepsilon = \int_{-\infty}^{\infty}h(\tau)\,\mathrm{d}\tau\int_{-\infty}^{\infty}h(\sigma)\,\mathrm{d}\sigma\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}f_i(t-\tau)f_i(t-\sigma)\,\mathrm{d}t - {} - 2\int_{-\infty}^{\infty}h(\tau)\,\mathrm{d}\tau\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}f_d(t)\,f_i(t-\tau)\,\mathrm{d}t + \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}f_d^2(t)\,\mathrm{d}t\,.$$

Since by (1)

$$(15)\qquad \varphi_{ii}(\tau-\sigma) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}f_i(t-\tau)\,f_i(t-\sigma)\,\mathrm{d}t$$

is the autocorrelation function of the input with the argument $(\tau-\sigma)$, and by (4) and (5)

$$(16)\qquad \varphi_{id}(\tau) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}f_d(t)f_i(t-\tau)\,\mathrm{d}t$$

is the input–desired output crosscorrelation function, and

$$(17)\qquad \varphi_{dd}(0) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}f_d^2(t)\,\mathrm{d}t$$

is the mean-square value of the desired output expressed as the value of the autocorrelation function of the desired output for $\tau=0$, we write (14) in terms of (15)–(17) and find that

$$(18)\qquad \varepsilon = \int_{-\infty}^{\infty}h(\tau)\,\mathrm{d}\tau\int_{-\infty}^{\infty}h(\sigma)\,\mathrm{d}\sigma\varphi_{ii}(\tau-\sigma) - 2\int_{-\infty}^{\infty}h(\tau)\,\mathrm{d}\tau\varphi_{id}(\tau) + \varphi_{dd}(0)\,.$$

This is the relationship we wanted to establish. Obviously, the mean-square error is dependent upon the system characteristic $h(t)$, but it is interesting and important to note that it is also dependent upon the autocorrelation function of the input $\varphi_{ii}(\tau)$ and the input–desired output crosscorrelation of the system $\varphi_{id}(\tau)$, as well as the mean-square value of the desired output $\varphi_{dd}(0)$. Let us remember that the mean-square error (18) holds for any given $h(t)$, not necessarily the one that minimizes the mean-square error, any given $\varphi_{ii}(\tau)$ and any given $\varphi_{id}(\tau)$. In other words, for any given system, characterized by $h(t)$, with an input $f_i(t)$ which has an autocorrelation function $\varphi_{ii}(\tau)$, and a desired output $f_d(t)$ which has the crosscorrelation function $\varphi_{id}(\tau)$ with $f_i(t)$, the mean square error as defined by (11) is expressed by (18).

The next step in the problem of filtering and prediction is the determination of a condition under which the mean-square error (18) is the minimum. From the observations we have just made it is clear that the condition will involve $\varphi_{ii}(\tau)$ and $\varphi_{id}(\tau)$, the two functions which specify the problem.

3. – Minimization of mean-square error.

Since in filtering and prediction, $f_i(t)$ and $f_d(t)$ are specified for any particular problem so that $\varphi_{ii}(\tau)$ and $\varphi_{id}(\tau)$ are completely determined at the outset, the only change that can be made for reducing the mean-square error is a change in the system characteristic $h(t)$. The finding of a condition relating $h(t)$, $\varphi_{ii}(\tau)$ and $\varphi_{id}(\tau)$ for minimum meansquare error is a problem in the calculus of variations. Accordingly we let $\in$ be a parameter which is independent of $h(t)$, and $\eta(t)$ be the variation of $h(t)$ with the condition that

$$\eta(t) = 0 \qquad \text{for } t < 0. \tag{19}$$

Since $h(t)$ is the system time response to a unit-impulse excitation applied at $t = 0$, and the system is initially at rest, it is necessary that

$$h(t) = 0 \qquad \text{for } t < 0. \tag{20}$$

It follows immediately that if $\eta(t)$ is a possible variation of $h(t)$ it must satisfy the condition (19). When $h(t)$ undergoes the variation $\in\eta(t)$ we let the corresponding variation in ε be $\delta\varepsilon$. It is known that a necessary condition for minimum mean square error, $\varepsilon_{\min}$ is that

$$\frac{\partial}{\partial \in}(\varepsilon + \delta\varepsilon)\,|\in = 0 \qquad \text{for all possible } \eta\,. \tag{21}$$

Introducing the variations in (18) we have

$$(22)\qquad \varepsilon + \delta\varepsilon = \int_{-\infty}^{\infty} [h(\tau) + \epsilon\eta(\tau)]\,\mathrm{d}\tau \int_{-\infty}^{\infty} [h(\sigma) + \epsilon\eta(\sigma)]\,\mathrm{d}\sigma\,\varphi_{ii}(\tau - \sigma) - \\ - 2\int_{-\infty}^{\infty} [h(\tau) + \epsilon\eta(\tau)]\,\mathrm{d}\tau\,\varphi_{id}(\tau) + \varphi_{dd}(0)\,,$$

which can be simplified to

$$(23)\qquad \varepsilon + \delta\varepsilon = \varepsilon + 2\epsilon \int_{-\infty}^{\infty} \eta(\tau)\,\mathrm{d}\tau \int_{-\infty}^{\infty} h(\sigma)\,\mathrm{d}\sigma\,\varphi_{ii}(\tau - \sigma) + \\ + \epsilon^2 \int_{-\infty}^{\infty} \eta(\tau)\,\mathrm{d}\tau \int_{-\infty}^{\infty} \eta(\sigma)\,\mathrm{d}\sigma\,\varphi_{ii}(\tau - \sigma) - 2\epsilon \int_{-\infty}^{\infty} \eta(\tau)\,\mathrm{d}\tau\,\varphi_{id}(\tau)\,.$$

Applying condition (21) to (23) we find that for $\varepsilon_{\min}$ it is necessary that

$$(24)\qquad \int_{-\infty}^{\infty} \eta(\tau)\,\mathrm{d}\tau \int_{-\infty}^{\infty} h(\sigma)\,\mathrm{d}\sigma\,\varphi_{ii}(\tau - \sigma) - \int_{-\infty}^{\infty} \eta(\tau)\,\mathrm{d}\tau\,\varphi_{id}(\tau) = 0 \qquad \text{for all possible } \eta\,.$$

Rewriting (24) we have

$$(25)\qquad \int_{-\infty}^{\infty} \eta(\tau)\,\mathrm{d}\tau \left[\int_{-\infty}^{\infty} h(\sigma)\varphi_{ii}(\tau - \sigma)\mathrm{d}\sigma - \varphi_{id}(\tau)\right] = 0 \qquad \text{for all possible } \eta\,.$$

Since the range for τ is $(-\infty, \infty)$ we shall consider (25) for $(-\infty < \tau < 0)$ and then for $(0 \leqslant \tau < \infty)$. According to (19), $\eta(\tau) = 0$ for $(-\infty < \tau < 0)$. Hence (25) holds although

$$(26)\qquad \int_{-\infty}^{\infty} h(\sigma)\varphi_{ii}(\tau - \sigma)\,\mathrm{d}\sigma - \varphi_{id}(\tau)\,,$$

is not necessarily zero, and it generally does not vanish for $(-\infty < \tau < 0)$. On the other hand, for $(0 \leqslant \tau < 0)$, the expression (26) in (25) must vanish because if it does not, an $\eta(\tau)$ can be found such that the condition (25) is violated. The conclusion is that a necessary condition for $\varepsilon_{\min}$ is that

$$(27)\qquad \int_{-\infty}^{\infty} h_{\mathrm{opt}}(\sigma)\varphi_{ii}(\tau - \sigma)\,\mathrm{d}\sigma - \varphi_{id}(\tau) = 0 \qquad \text{for} \quad \tau \geqslant 0\,.$$

Here we let $h_{\text{opt}}(t)$ denote the time response of the optimum linear system to a unit-impulse excitation. We call a system an optimum system when the mean-square error is the minimum. Equation (27) is of the Wiener-Hopf type, and we shall call it the Wiener-Hopf equation.

Since (21) is also a condition for maximum mean-square error, we should establish the fact that (27) is for minimum not maximum mean-square error. This fact can be shown quite readily, but for a short presentation of the subject the details will be omitted.

4. – Solution of the Wiener-Hopf equation.

Before the discussion on the solution of the Wiener-Hopf equation which will yield the optimum linear system characteristic in terms of the transforms of the input autocorrelation function and the input-desired output crosscorrelation function, we should consider some background material.

4·1. *Transforms of correlation functions.* – An extremely important theorem in the theory of filtering and prediction is the Wiener theorem for autocorrelation. It states that for a stationary random process $f_a(t)$ whose autocorrelation function is

$$\varphi_{aa}(\tau) = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} f_a(t) f_a(t+\tau)\,\mathrm{d}t\,, \tag{28}$$

the power density spectrum $\Phi_{aa}(\omega)$ of $f_a(t)$ and its autocorrelation function are Fourier transforms of each other, that is,

$$\varphi_{aa}(\tau) = \int_{-\infty}^{\infty} \Phi_{aa}(\omega) \exp[j\omega\tau]\,\mathrm{d}\omega\,, \tag{29}$$

and

$$\Phi_{aa}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi_{aa}(\tau) \exp[-j\omega\tau]\,\mathrm{d}\tau\,, \tag{30}$$

where ω is the angular frequency.

For the physical meaning of the power density spectrum let us consider the wave of Fig. 1 whose autocorrelation function is given by (3). In accordance with the Wiener theorem, the power density spectrum of the wave is

$$\Phi_{aa}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} E^2 \exp[-2k|\tau|] \exp[-j\omega\tau]\,\mathrm{d}\tau = \frac{E^2}{\pi}\,\frac{2k}{(2k)^2+\omega^2}\,. \tag{31}$$

A sketch of this function is given in Fig. 3. If we consider the wave in Fig. 1 as a voltage or a current associated with a one-ohm load of pure resistance then the sum of the shaded areas under the curve of $\Phi_{aa}(\omega)$ between the bands (ω_1, ω_2) and $(-\omega_1, -\omega_2)$ represents the power supplied to the load by components of $f_a(t)$ of all frequencies in the band (ω_1, ω_2). The total area under the curve represents the total power supplied to the load since

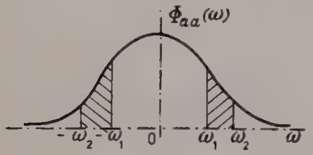


Fig. 3.

$$(32)\qquad \int\limits_{-\infty}^{\infty}\Phi_{aa}(\omega)\,\mathrm{d}\omega = \varphi_{aa}(0) = \lim_{T\to\infty}\frac{1}{2T}\int\limits_{-T}^{T} f_a^2(t)\,\mathrm{d}t\,.$$

Analogous to the Wiener theorem we can relate the crosscorrelation function $\varphi_{ab}(\tau)$, for $f_a(t)$ and $f_b(t)$, defined as

$$(33)\qquad \varphi_{ab}(\tau) = \lim_{T\to\infty}\frac{1}{2T}\int\limits_{-T}^{T} f_a(t)\,f_b(t+\tau)\,\mathrm{d}t\,,$$

to its Fourier transform $\Phi_{ab}(\omega)$ by the expressions

$$(34)\qquad \varphi_{ab}(\tau) = \int\limits_{-\infty}^{\infty}\Phi_{ab}(\omega)\exp[j\omega\tau]\,\mathrm{d}\omega\,,$$

and

$$(35)\qquad \Phi_{ab}(\omega) = \frac{1}{2\pi}\int\limits_{-\infty}^{\infty}\varphi_{ab}(\tau)\exp[-j\omega\tau]\,\mathrm{d}\tau\,.$$

We shall call $\Phi_{ab}(\omega)$ the cross-power density spectrum of $f_a(t)$ and $f_b(t)$. Although it possible to consider a particular situation where $f_a(t)$, $f_b(t)$ and $\Phi_{ab}(\omega)$ correspond to physical quantities, we find it unnecessary to look for a physical interpretation of $\Phi_{ab}(\omega)$ in every problem. Unless it has a meaningful and direct physical interpretation in a problem we shall consider the cross-power density as a mathematical quantity.

4·2. *The input-output crosscorrelation theorem for a linear system.* – For a linear system with the unit-impulse response $h(t)$, an input $f_i(t)$ which is a stationary random process, and the corresponding output $f_o(t)$, an important relationship exists among $h(t)$, the input autocorrelation $\varphi_{ii}(\tau)$ and the input-output crosscorrelation $\varphi_{io}(\tau)$. To derive this relationship let us first write the expression for the input-output crosscorrelation of the linear system according to definition. Thus

$$\varphi_{io}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f_i(t) f_o(t+\tau)\,\mathrm{d}t\,. \tag{36}$$

To bring the system characteristic into this expression we introduce the convolution integral (6). In so doing we have

$$\varphi_{io}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f_i(t)\,\mathrm{d}t \int_{-\infty}^{\infty} h(\sigma) f_i(t+\tau-\sigma)\,\mathrm{d}\sigma\,. \tag{37}$$

By inversion of the order of integration, (37) becomes

$$\varphi_{io}(\tau) = \int_{-\infty}^{\infty} h(\sigma)\,\mathrm{d}\sigma \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f_i(t) f_i(t+\tau-\sigma)\,\mathrm{d}t\,. \tag{38}$$

Since

$$\varphi_{ii}(\tau-\sigma) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f_i(t) f_i(t+\tau-\sigma)\,\mathrm{d}t\,, \tag{39}$$

it follows that (38) is

$$\varphi_{io}(\tau) = \int_{-\infty}^{\infty} h(\sigma) \varphi_{ii}(\tau-\sigma)\,\mathrm{d}\sigma\,. \tag{40}$$

We have therefore shown the following theorem: The input-output crosscorrelation of a linear system is the convolution of the system response to unit-impulse excitation and the input autocorrelation.

This theorem plays an important role in the statistical theory of communication.

4·3. *Relation between the unit-impulse response and the system function.* – In the frequency domain a linear system may be characterized by its function $H(\omega)$ which is the ratio of the output to the input, as a function of the angular frequency ω, when the input is a steady sinusoidal voltage or current. As we know, the input and output are expressed in the complex form so that $H(\omega)$ is a complex expression which contains the amplitude and phase spectrums of the system. We shall state here for reference the well known relations between the linear system unit-impulse response $h(t)$ and the system function. These relations are:

$$h(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H(\omega) \exp[j\omega t]\,\mathrm{d}\omega\,, \tag{41}$$

and

$$H(\omega) = \int_{\infty}^{\infty} h(t) \exp\left[-i\omega t\right] \mathrm{d}t \; . \tag{42}$$

Now, having introduced the necessary background material in Sect. 4·1 to 4·3, we return to the Wiener-Hopf equation (27) for discussion and solution. Noting that there is great similarity in appearance between (27) and (40) we shall investigate the significance of the Wiener-Hopf equation particularly with respect to the restriction $\tau \geqslant 0$.

Let us first consider the term

$$\int_{-\infty}^{\infty} h_{\text{opt}}(\sigma)\, \varphi_{ii}(\tau - \sigma)\, \mathrm{d}\sigma \, , \tag{43}$$

in the Wiener-Hopf equation.

A sketch of the functions involved, just for the purpose of explanation without any reference to a specific problem and disregarding precision in drawing is given in Fig. 4(*a*).

The convolution of $h_{\text{opt}}(\sigma)$ and $\varphi_{ii}(\sigma)$ results in the curve of Fig. 4(*b*). In the light of the input-output crosscorrelation theorem (40) the convolution (43) for all values of τ must be equal to the input-output crosscorrelation of the optimum system. Now, if we refer to (27) and consider the whole statement we find that it states that the optimum linear system must be such that its input-output crosscorrelation is equal to the input-desired output crosscorrelation for $\tau \geqslant 0$. It must be emphasized that for $\tau < 0$ these crosscorrelations need not be equal, and are generally not equal. The reason for this is that the desired output is generally not possible to obtain without error under the conditions of the problem. If the desired output were possible to obtain without error then (27) holds for all values of τ and the solution of the problem is trivial. The difference between (43) and $\varphi_{id}(\tau)$ which is sketched in Fig. 4(*c*), in accordance with the Wiener-Hopf equation is therefore zero for

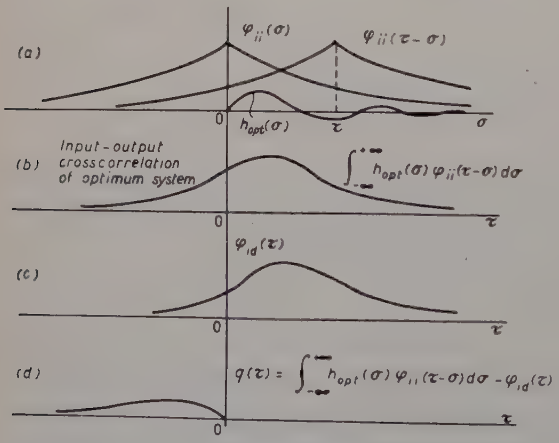

(a)
$\varphi_{ii}(\sigma)$
$\varphi_{ii}(\tau-\sigma)$
0
$h_{opt}(\sigma)$
τ
σ
(b)
Input-output crosscorrelation of optimum system
$\int_{-\infty}^{+\infty} h_{opt}(\sigma)\,\varphi_{ii}(\tau-\sigma)\,d\sigma$
0
τ
(c)
$\varphi_{id}(\tau)$
0
τ
(d)
$q(\tau) = \int_{-\infty}^{+\infty} h_{opt}(\sigma)\,\varphi_{ii}(\tau-\sigma)\,d\sigma - \varphi_{id}(\tau)$
0
τ

Fig. 4.

$\tau \geqslant 0$, and is generally not zero for $\tau < 0$ as illustrated in Fig. 4(d). Let us put

$$q(\tau) = \int_{-\infty}^{\infty} h_{\text{opt}}(\sigma)\varphi_{ii}(\tau - \sigma)\,\mathrm{d}\sigma - \varphi_{id}(\tau)\,. \tag{44}$$

This consideration of the Wiener-Hopf equation shows that to obtain minimum mean-square error the linear system must be so designed that its input-output crosscorrelation equals the input-output crosscorrelation for $\tau \geqslant 0$. For all other values of τ, that is, for $\tau < 0$, there is no restriction.

The solution of the Wiener Hopf equation starts from the fact that $q(\tau)$ as given by (44) is a function that vanishes for $\tau \geqslant 0$. For this reason let us note that if a function $f(t) \to 0$ as $t \to \infty$ and $f(t) = 0$ for $t < 0$, as illustrated in Fig. 5(a), then its Fourier transform $F(\lambda)$ as a function of the complex variable $\lambda = \omega + j\sigma$ has no pole in the lower half-plane (lhp) as shown in Fig. 5(b). The poles of $F(\lambda)$ are in the upper half-plane (uhp). We have

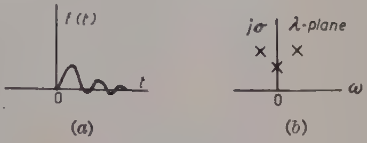


Fig. 5.

$$F(\lambda) = \frac{1}{2\pi}\int_{0}^{\infty} f(t)\exp[-j\lambda t]\,\mathrm{d}t\,, \tag{45}$$

and

$$f(t) = \int_{-\infty}^{\infty} F(\lambda)\exp[i\lambda t]\,\mathrm{d}\lambda\,. \tag{46}$$

As an example, consider

$$f(t) = \begin{cases} A\exp[-at] & \text{for} \quad t > 0\,, \\ 0 & \text{for} \quad t < 0\,, \end{cases} \tag{47}$$

then

$$F(\lambda) = \frac{1}{2\pi}\int_{0}^{\infty} A\exp[-at]\exp[-j\lambda t]\,\mathrm{d}t = \frac{A}{2\pi}\frac{1}{a + j\lambda}\,, \tag{48}$$

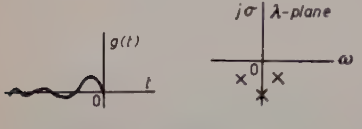
and the pole of $F(\lambda)$ is found from

$$a + j\lambda = 0\,, \tag{49}$$

or

$$a + j(\omega + j\sigma) = 0\,, \tag{50}$$

which gives the location of the pole in the uhp as



(a) Fig. 6. (b)

$$\omega = 0\,, \qquad j\sigma = ja\,. \tag{51}$$

On the other hand if $g(t) \to 0$ as $t \to -\infty$ and $g(t) = 0$ for $t > 0$, as illustrated in Fig. 6(a), then its Fourier transform $G(\lambda)$ as a function of λ has no poles in the uhp. The poles of $G(\lambda)$ are in the lhp, as shown in Fig. 6(b). The relations between $g(t)$ and $G(\lambda)$ are

$$G(\lambda) = \frac{1}{2\pi}\int_{-\infty}^{\infty} g(t) \exp[-j\lambda t]\,\mathrm{d}t\,, \tag{52}$$

and

$$g(t) = \int_{-\infty}^{\infty} G(\lambda) \exp[j\lambda t]\,\mathrm{d}\lambda\,. \tag{53}$$

Returning to $q(\tau)$ in (44) we see that it behaves as $g(t)$ so that its Fourier transform has no poles in the uhp. Let

$$Q(\lambda) = \frac{1}{2\pi}\int_{-\infty}^{\infty} q(\tau) \exp[-j\lambda\tau]\,\mathrm{d}\tau\,. \tag{54}$$

Then

$$Q(\lambda) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp[-j\lambda\tau]\,\mathrm{d}\tau\left[\int_{-\infty}^{\infty} h_{\mathrm{opt}}(\sigma)\,\varphi_{ii}(\tau-\sigma)\,\mathrm{d}\sigma - \varphi_{id}(\tau)\right] = \tag{55}$$

$$= H_{\mathrm{opt}}(\lambda)\,\Phi_{ii}(\lambda) - \Phi_{id}(\lambda)\,,$$

in which $H_{\mathrm{opt}}(\lambda)$ is the system function of the optimum system, as a function of λ, which is related to $h_{\mathrm{opt}}(t)$ by (42). The input power density spectrum $\Phi_{ii}(\lambda)$ and the input-desired output cross-power density spectrum $\Phi_{id}(\lambda)$ related to $\varphi_{ii}(\tau)$ and $\varphi_{id}(\tau)$ by (30) and (35) respectively. Because of the fact that $q(\tau)$ vanishes for $\tau > 0$, we have the result that $Q(\lambda)$, or

$$[H_{\mathrm{opt}}(\lambda)\,\Phi_{ii}(\lambda) - \Phi_{id}(\lambda)] \quad \text{has no poles in the uhp}\,. \tag{56}$$

At this point it is necessary to introduce the idea of spectrum factorization. Let us start with an example. If the input is the sum of a message and a noise then

$$f_i(t) = f_m(t) + f_n(t) \tag{57}$$

and the input autocorrelation is

$$\varphi_{ii}(\tau) = \varphi_{mm}(\tau) + \varphi_{mn}(\tau) + \varphi_{nn}(\tau) + \varphi_{nm}(\tau)\,. \tag{58}$$

After transformation in accordance with (30) and (35) we have

$$\Phi_{ii}(\omega) = \Phi_{mm}(\omega) + \Phi_{nn}(\omega) + \Phi_{mn}(\omega) + \Phi_{nm}(\omega)\,. \tag{59}$$

In solving the Wiener-Hopf equation by the method of spectrum factorization we need only factorize the input density spectrum. To show the idea let us assume that the message $f_m(t)$ is a Poisson rectangular wave whose power density spectrum is given by (31). Furthermore, for simplicity, we put

$$\Phi_{mm}(\omega) = \frac{1}{1+\omega^2}\,. \tag{60}$$

If the noise $f_n(t)$ is a white noise, then

$$\Phi_{nn}(\omega) = a^2\,, \tag{61}$$

which is a constant. We further assume that $\Phi_{mn}(\omega) = 0$. The input power density spectrum (59), for the case of (57), becomes

$$\Phi_{ii}(\omega) = \frac{1}{1+\omega^2} + a^2 = \frac{1+a^2+a^2\omega^2}{1+\omega^2}\,. \tag{62}$$

Writing this expression in λ and factorizing we obtain

$$\Phi_{ii}(\lambda) = \frac{a(b+j\lambda)}{1+j\lambda}\,\frac{a(b-j\lambda)}{1-j\lambda}\,, \tag{63}$$

where $b = \sqrt{1+a^2}/a$. If we put

$$\Phi_{ii}^{+} = \frac{a(b+j\lambda)}{1+j\lambda}\,, \tag{64}$$

we shall find that $\Phi_{ii}^{+}(\lambda)$ has a pole in the uhp at $(\omega = 0,\ j\sigma = j1)$ and a zero in the uhp at $(\omega = 0,\ j\sigma = jb)$ as shown in Fig. 7(a).

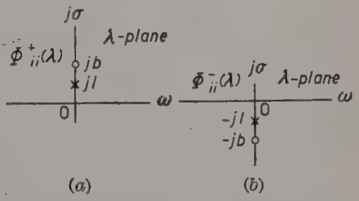


Fig. 7.

Similarly we put

$$\Phi_{ii}^{-}(\lambda) = \frac{a(b-j\lambda)}{1-j\lambda}\,, \tag{65}$$

which has a pole in the lhp at $(\omega = 0,\ j\sigma = -j1)$ and a zero in the lhp at $(\omega = 0,\ j\sigma = -jb)$ as shown in Fig. 7(*b*).

In the factorization of the input power density spectrum we assume that it is possible to express the spectrum in the form

$$\Phi_{ii}(\lambda) = \Phi_{ii}^{+}(\lambda)\Phi_{ii}^{-}(\lambda)\,, \tag{66}$$

in which $\Phi_{ii}^{+}(\lambda)$ contains all the poles and zeros of $\Phi_{ii}(\lambda)$ that are in the uhp and $\Phi_{ii}(\lambda)$ contains all the poles and zeros of $\Phi_{ii}(\lambda)$ that are in the lhp. Furthermore

$$\Phi_{ii}^{+}(\lambda) = \overline{\Phi_{ii}^{-}(\lambda)}\,, \tag{67}$$

where the bar over $\Phi_{ii}^{-}(\lambda)$ denotes the conjugate of the function. It is important to note that $\Phi_{ii}(\omega)$ is an even function, and that if it is expressed as a rational function we shall be able to factorize it in accordance with (66) and (67).

We now return to the consideration of (56). If we multiply this expression by $1/\Phi_{ii}^{-}(\lambda)$ we shall find that the resulting expression still has no poles in the uhp. To see this let us note that $\Phi_{ii}^{-}(\lambda)$ has no zeros in the uhp so that its reciprocal cannot have poles in the plane. The result of the multiplication is that

$$\left[H_{\text{opt}}(\lambda)\,\Phi^{+}(\lambda) - \frac{\Phi_{id}(\lambda)}{\Phi_{ii}^{-}(\lambda)}\right] \qquad \text{has no poles in uhp}\,. \tag{68}$$

For simplicity in the writing of the next few steps we put

$$\Psi(\lambda) = \frac{\Phi_{id}(\lambda)}{\Phi_{ii}^{-}(\lambda)}\,, \tag{69}$$

$$\psi(t) = \int_{\infty}^{\infty} \Psi(\lambda) \exp[j\lambda t]\,\mathrm{d}\lambda \tag{70}$$

and

$$\Psi(\lambda) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \psi(t) \exp[-j\lambda t]\,\mathrm{d}t\,. \tag{71}$$

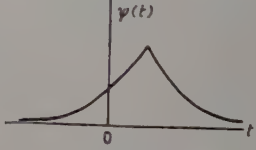


Fig. 8.

Since $\Phi_{id}(\lambda)/\Phi_{ii}^{-}(\lambda)$ in general has poles in both half planes, its transform $\psi(t)$ in general does not vanish over a half line. An example of $\psi(t)$ is shown in Fig. 8. We shall write the right hand side of (71) as the sum of two integrals, one over the interval $(-\infty, 0)$ and the other over the interval $(0, \infty)$. Thus

$$\Psi(\lambda) = \frac{1}{2\pi}\int_{-\infty}^{0}\psi(t)\exp[-j\lambda t]\,dt + \frac{1}{2\pi}\int_{0}^{\infty}\psi(t)\exp[-j\lambda t]\,dt\,. \tag{72}$$

Substituting (72) in (68) we have

$$\left[H_{\text{opt}}(\lambda)\,\Phi_{ii}^{+}(\lambda) - \frac{1}{2\pi}\int_{0}^{\infty}\psi(t)\exp[-j\lambda t]\,dt - \right. \\ \left. - \frac{1}{2\pi}\int_{-\infty}^{0}\psi(t)\exp[-j\lambda t]\,dt\right] \quad \text{has no poles in the uhp}\,. \tag{73}$$

By considering the location of the poles in the expression as a whole, and in the components we shall finally obtain the solution of the Wiener-Hopf equation. First, the optimum system function $H_{\text{opt}}(\lambda)$ must have no poles in the lhp because its transform, the response to unit-impulse excitation, behaves as $f(t)$ in Fig. 5(a). Next, the function $\Phi_{ii}(\lambda)$ has no poles in the lhp by definition. Hence

$$H_{\text{opt}}(\lambda)\,\Phi_{ii}^{+}(\lambda) \qquad \text{has no poles in the lhp}\,. \tag{74}$$

The term

$$\frac{1}{2\pi}\int_{0}^{\infty}\psi(t)\exp[-j\lambda t]\,dt \qquad \text{has no poles in the lhp}\,, \tag{75}$$

because it is the transform of a function that behaves as $f(t)$ in Fig. 5(a). Therefore we see that the first two terms of (73)

$$\left[H_{\text{opt}}(\lambda)\Phi_{ii}^{+}(\lambda) - \frac{1}{2\pi}\int_{0}^{\infty}\psi(t)\exp[-j\lambda t]\,dt\right] \qquad \text{has no poles in the lhp}\,. \tag{76}$$

The property of (73), that it has no poles in the uhp, must now be utilized. In (73) the term

$$\frac{1}{2\pi}\int_{-\infty}^{0}\psi(t)\exp[-j\lambda t]\,dt \qquad \text{has no poles in the uhp}\,,$$

because this is the transform of a function that behaves as $g(t)$ in Fig. 6(a). The location of poles of this term is in agreement with the overall property of (73). To satisfy this overall property it is therefore necessary that the remainder of the expression

$$(77) \qquad \left[H_{\text{opt}}(\lambda)\,\Phi^{+}_{ii}(\lambda) - \frac{1}{2\pi}\int\limits_{0}^{\infty}\psi(t)\exp\left[-j\lambda t\right]\mathrm{d}t\right] \qquad \text{has no poles in the uhp}\,.$$

Now, since this expression has been found to have no poles in the lhp also, as stated in (76), it is an expression that has no poles in the whole plane. We conclude that

$$(78) \qquad H_{\text{opt}}(\lambda)\,\Phi^{+}_{ii}(\lambda) - \frac{1}{2\pi}\int\limits_{0}^{\infty}\psi(t)\exp\left[-i\lambda t\right]\mathrm{d}t = \text{const}\,.$$

We can show that the constant should be zero by further analysis, but we shall omit the details here. We have, finally, from (78)

$$(79) \qquad H_{\text{opt}}(\lambda) = \frac{1}{2\pi\Phi^{+}_{ii}(\lambda)}\int\limits_{0}^{\infty}\psi(t)\exp\left[-j\lambda t\right]\mathrm{d}t\,,$$

where

$$(80) \qquad \psi(t) = \int\limits_{-\infty}^{\infty}\frac{\Phi_{id}(\lambda)}{\Phi^{-}_{ii}(\lambda)}\exp\left[i\lambda t\right]\mathrm{d}\lambda\,.$$

This is the solution of the Wiener-Hopf equation (27). It gives the optimum system function explicitly in terms of the input power density spectrum in its factorized form, and the input-desired output cross-power density spectrum. We must emphasize the fact that the solution is in a general form without restriction on the manner in which messages and noise are combined, and without restriction on the form of the desired output as long as it has a nonzero correlation with the input. Obviously if the desired output and the input have a zero correlation the problem is a trivial one. A common form of the input is the sum of a message and a noise, but the theory is not restricted to this form of input. However, we must bear in mind that in some problems the optimum system may yield a very poor result because the theory is restricted to the linear system. A case in point is that in which the input is the product of a message and a noise, and the desired output is the message. Let us again observe that given a stationary random input with a power density spectrum $\Phi_{ii}(\omega)$ that can be factorized in accordance with (66), and a stationary random desired output that has a cross-power density spectrum $\Phi_{id}(\omega)$ with the input, then (79) and (80) specify the optimum linear system in terms of these spectrums. This is a very general and important result which applies to a large class of problems.

5. – Illustrative examples of filtering and prediction.

5·1. *Filtering.* – Referring to the discussion under Sect. **2**, if the input in a filter is

$$f_i(t) = f_m(t) + f_n(t) \tag{81}$$

then for a lag filter we specify that

$$f_d(t) = f_m(t - \alpha) \tag{82}$$

in which $\alpha \geqslant 0$ is the lag time, and for a lead (prediction) filter we specify that

$$f_d(t) = f_m(t + \alpha) \tag{83}$$

in which $\alpha \geqslant 0$ is the prediction time.

In applying (79) and (80) to the present problem we first determine $\varphi_{id}(\tau)$. Accordingly we have

$$\varphi_{id}(\tau) = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} f_i(t) f_d(t + \tau)\,\mathrm{d}t = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} f_i(t) f_m(t+\tau\pm\alpha)\,\mathrm{d}t = \varphi_{im}(\tau\pm\alpha)\,. \tag{84}$$

From this crosscorrelation function we obtain $\Phi_{id}(\lambda)$ by transformation thus

$$\Phi_{id}(\lambda) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi_{im}(\tau \pm \alpha) \exp[-j\lambda\tau]\,\mathrm{d}\tau = \exp[\pm j\alpha\lambda]\,\Phi_{im}(\lambda)\,. \tag{85}$$

Substituting this result in (80) and retaining the general form (79) we have the following formulas for the lag filter or the lead filter:

$$H_{\text{opt}}(\lambda) = \frac{1}{2\pi\Phi^{+}_{ii}(\lambda)}\int_{0}^{\infty}\psi(t)\exp[-j\lambda t]\,\mathrm{d}t\,, \tag{86}$$

$$\psi(t) = \int_{-\infty}^{\infty}\frac{\Phi_{im}(\lambda)}{\Phi^{-}_{ii}(\lambda)}\exp[j(t\pm\alpha)\lambda]\,\mathrm{d}\lambda\,, \tag{87}$$

where $-\alpha$ is for the lag filter and $+\alpha$ is for the lead filter. In (86) $\Phi_{ii}(\lambda)$ is given by (59) and in (87)

$$\Phi_{im}(\lambda) = \Phi_{mm}(\lambda) + \Phi_{nm}(\lambda)\,. \tag{88}$$

For a specific example, let us assume that $\Phi_{mm}(\omega) = 1/(1+\omega^2)$, $\Phi_{nn}(\omega) = a^2$, and $\Phi_{mn}(\omega) = 0$. This example has been chosen to illustrate spectrum factorization (see (60) to (65)). We shall determine the system function of the optimum lag filter. Substituting (65) in (87) and noting that $\Phi_{im}(\lambda) = \Phi_{mm}(\lambda)$, we find

$$\psi(t) = \int_{-\infty}^{\infty} \frac{1}{1+\lambda^2} \, \frac{1-j\lambda}{a(b-j\lambda)} \exp[j(t-\alpha)\lambda] \, \mathrm{d}\lambda = \tag{89}$$

$$= \begin{cases} \dfrac{2\pi}{a(1+b)} \exp[-(t-\alpha)] & \text{for} \quad t > \alpha\,, \\[2ex] \dfrac{2\pi}{a(1+b)} \exp[b(t-\alpha)] & \text{for} \quad t < \alpha\,. \end{cases}$$

We shall next evaluate the transform

$$\frac{1}{2\pi} \int_0^{\infty} \psi(t) \exp[-j\lambda t] \, \mathrm{d}t \tag{90}$$

in (86). This transform is

$$\frac{1}{2\pi} \int_0^{\infty} \psi(t) \exp[-j\lambda t] \, \mathrm{d}t = \frac{1}{a(1+b)} \left| \int_0^{\alpha} \exp[b(t-\alpha)] \exp[-j\lambda t] + \right. \tag{91}$$

$$\left. + \int_{\alpha}^{\infty} \exp[-(t-\alpha)] \exp[-j\lambda t] \, \mathrm{d}t \right| =$$

$$= \frac{1}{a(1+b)} \, \frac{1}{(b-j\lambda)(1+j\lambda)} \left[(1+b) \exp[-j\alpha\lambda] - (1+j\lambda) \exp[-b\alpha]\right].$$

Finally, according to (86), we multiply this expression by $1/\Phi_{ii}^{+}(\lambda)$ to obtain the optimum system function. The result is that

$$H_{\text{opt}}(\lambda) = \frac{1}{a^2(1+b)} \, \frac{1}{b^2+\lambda^2} \left[(1+b)\exp[-j\alpha\lambda] - (1+j\lambda)\exp[-b\alpha]\right], \tag{92}$$

characterizes the optimum lag filter.

The specification of the optimum filter characteristic in the form (92) may be inconvenient for certain problems, such as synthesis. In such a situation it may be helpful to specify the filter in the time domain. Applying (41) to (92) we find that the optimum lag filter has the unit-impulse response

$$h_{\text{opt}}(t) = \begin{cases} 0 & \text{for} \quad -\infty < t < 0 \\ \dfrac{\exp[-b\alpha]}{a^2b(1+b)}[b\cosh bt + \sinh bt] & \text{for} \quad 0 < t < \alpha \\ \dfrac{\exp[-bt]}{a^2b(1+b)}[b\cosh b\alpha + \sinh b\alpha] & \text{for} \quad \alpha < t < \infty\,. \end{cases} \tag{93}$$

This response is sketched in Fig. 9.

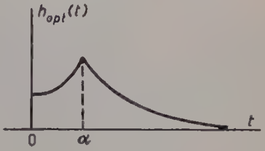


Fig. 9.

5·2. *Pure prediction.* – Referring to Sect. **2**, equations (9) and (10), we have for pure prediction

$$\varphi_{id}(\tau) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} f_i(t)\,f_d(t+\tau)\,\mathrm{d}t = \tag{94}$$

$$= \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} f_m(t)f_m(t+\tau+\alpha)\,\mathrm{d}t = \varphi_{mm}(\tau+\alpha)\,.$$

The transform of this function is

$$\Phi_{id}(\lambda) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\varphi_{mm}(\tau+\alpha)\exp[-j\lambda\tau]\,\mathrm{d}\tau = \exp[j\alpha\lambda]\,\Phi_{mm}(\lambda)\,. \tag{95}$$

Since $f_i(t) = f_m(t)$ we have

$$\Phi_{ii}(\lambda) = \Phi_{mm}(\lambda) = \Phi^{+}_{mm}(\lambda)\,\Phi^{-}_{mm}(\lambda) \tag{96}$$

so that

$$\Phi^{+}_{ii}(\lambda) = \Phi^{+}_{mm}(\lambda) \tag{97}$$

and

$$\Phi^{-}_{ii}(\lambda) = \Phi^{-}_{mm}(\lambda)\,. \tag{98}$$

The ratio $\Phi_{id}(\lambda)/\Phi^{-}_{ii}(\lambda)$ in (80) is therefore

$$\frac{\Phi_{id}(\lambda)}{\Phi^{-}_{ii}(\lambda)} = \frac{\Phi_{mm}(\lambda)}{\Phi^{-}_{mm}(\lambda)}\exp[j\alpha\lambda] = \Phi^{+}_{mm}(\lambda)\,\exp[j\alpha\lambda]\,. \tag{99}$$

With (97) and (99) we now write the formulas for the pure predictor, in accordance with the general expressions for the optimum linear system (79) and (80), as follows

$$H_{\mathrm{opt}}(\lambda) = \frac{1}{2\pi\Phi^{+}_{mm}(\lambda)} \int_{0}^{\infty} \psi(t) \exp\left[-j\lambda t\right] dt\,, \tag{100}$$

$$\psi(t) = \int_{-\infty}^{\infty} \Phi^{+}_{mm}(\lambda) \exp\left[j\lambda(t+\alpha)\right] d\lambda\,. \tag{101}$$

We note that pure prediction depends solely upon $\Phi^{+}_{mm}(\lambda)$.

For a specify example of pure prediction consider a message $f_m(t)$ whose power density spectrum is

$$\Phi_{mm}(\omega) = \frac{1}{(1+\omega^2)^2}\,. \tag{102}$$

For this spectrum we find

$$\Phi^{+}_{mm}(\lambda) = \frac{1}{(1+j\lambda)^2}\,. \tag{103}$$

Transforming this expression according to (101) we have

$$\begin{aligned} \psi(t) &= \int_{-\infty}^{\infty} \frac{1}{(1+j\lambda)^2} \exp\left[j\lambda(t+\alpha)\right] d\lambda \\ &= \begin{cases} 2\pi(t+\alpha)\exp\left[-(t+\alpha)\right] & \text{for} \quad t > -\alpha\,, \\ 0 & \text{for} \quad t < -\alpha\,. \end{cases} \end{aligned} \tag{104}$$

The transform in (100) is

$$\begin{aligned} \frac{1}{2\pi}\int_{0}^{\infty} \psi(t)\exp\left[-j\lambda t\right] dt &= \int_{0}^{\infty} (t+\alpha)\exp\left[-(t+\alpha)\right]\exp\left[-j\lambda t\right] dt = \\ &= \exp\left[-\alpha\right]\left[\frac{1}{(1+j\lambda)^2} + \frac{\alpha}{1+j\lambda}\right]. \end{aligned} \tag{105}$$

Therefore the system function of the optimum linear predictor for the message whose power density spectrum is given by (102) is

$$\begin{aligned} H_{\mathrm{opt}}(\lambda) &= (1+j\lambda)^2 \exp\left[-\alpha\right]\left[\frac{1}{(1+j\lambda)^2} + \frac{\alpha}{1+j\lambda}\right] = \\ &= \exp\left[-\alpha\right]\left[(1+\alpha) + j\alpha\lambda\right]. \end{aligned} \tag{106}$$

In this example it is easy to see that the system fnction (106) may be realized as in Fig. 10.

It is particularly interesting to note that the factor $\exp[-\alpha]$ indicates that whereas the output amplitude should be comparable with the input amplitude for very short prediction time, it should decrease exponentially as the prediction time increases. Accordingly, as the prediction time tends to infinity, the output should tend to zero. Practically, then, for very long prediction time the best output of the predictor on the mean-square criterion is zero output.

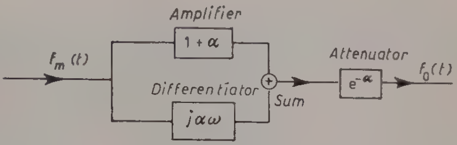


Fig. 10.

The amplification factor in the optimum system function is important, although practically when the wave form of a desired wave is satisfactory its amplitude is of secondary importance and sometimes it is of no consequence at all. The importance of the amplification factor can be seen from the fact that when the measure of error is the mean square error two waves of identical form will not have zero error unless their amplitudes are the same.

6. – Errors in filtering and prediction.

To find the expression for minimum mean square error in the theory of optimum systems we begin with the expression for means quare error (18). In this expression $h(t)$ is not necessarily the optimum system. However, when $h(t)$ satisfies the condition (27) the expression will then be for minimum mean square error. Therefore imposing the condition (27) in (18) we have the minimum mean square error,

$$\varepsilon_{\min} = \varphi_{mm}(0) - \int_{-\infty}^{\infty} h_{\text{opt}}(\tau)\varphi_{id}(\tau)\,\mathrm{d}\tau\,. \tag{107}$$

We shall consider the lag filter for which (86) and (87) have been derived. For this filter $\varphi_{dd}(0)=\varphi_{mm}(0)$, $\varphi_{id}(\tau)=\varphi_{im}(\tau-\alpha)$ so that

$$\varepsilon_{\min} = \varphi_{mm}(0) - \int_{-\infty}^{\infty} h_{\text{opt}}(\tau)\varphi_{im}(\tau-\alpha)\,\mathrm{d}\tau\,. \tag{108}$$

To reduce this expression to a form that shows the effect of the lag and involves simple computation, we shall substitute in (108) the transform (41) for $h_{\text{opt}}(\tau)$, that is, we shall put

$$(109) \qquad h_{\text{opt}}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} H_{\text{opt}}(\lambda)\exp[j\lambda\tau]\,\mathrm{d}\lambda\,.$$

Then for $H_{\text{opt}}(\lambda)$ we substitute the expression (86). In so doing we obtain

$$(110) \qquad \varepsilon_{\min} = \varphi_{mm}(0) - \int_{-\infty}^{\infty}\varphi_{im}(\tau-\alpha)\,\mathrm{d}\tau\,\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp[j\lambda\tau]\,\mathrm{d}\lambda\,\frac{1}{2\pi\Phi_{ii}^{+}(\lambda)}\cdot$$

$$\cdot\int_{0}^{\infty}\psi(t)\exp[-j\lambda t]\,\mathrm{d}t\,.$$

By inversion of the orders of integration we have

$$(111) \qquad \varepsilon_{\min} = \varphi_{mm}(0) - \frac{1}{2\pi}\int_{0}^{\infty}\psi(t)\,\mathrm{d}t\int_{-\infty}^{\infty}\frac{1}{\Phi_{ii}^{+}(\lambda)}\exp[-j\lambda t]\,\mathrm{d}\lambda\,\cdot$$

$$\cdot\frac{1}{2\pi}\int_{-\infty}^{\infty}\Phi_{im}(\tau-\alpha)\exp[j\lambda\tau]\,\mathrm{d}\tau\,.$$

With the change of variable $\tau-\alpha=v$ we get

$$(112) \qquad \varepsilon_{\min} = \varphi_{mm}(0) - \frac{1}{2\pi}\int_{0}^{\infty}\psi(t)\,\mathrm{d}t\int_{-\infty}^{\infty}\frac{1}{\Phi_{ii}^{+}(\lambda)}\exp[-j\lambda(t-\alpha)]\,\mathrm{d}\lambda\,\cdot$$

$$\cdot\frac{1}{2\pi}\int_{-\infty}^{\infty}\varphi_{im}(v)\exp[i\lambda v]\,\mathrm{d}v\,.$$

The integral on the extreme right-hand side of this equation is $\Phi_{im}^{-}(\lambda)$ so that (112) becomes

$$(113) \qquad \varepsilon_{\min} = \varphi_{mm}(0) - \frac{1}{2\pi}\int_{0}^{\infty}\psi(t)\,\mathrm{d}t\int_{-\infty}^{\infty}\frac{\Phi_{im}^{-}(\lambda)}{\Phi_{ii}^{+}(\lambda)}\exp[-j\lambda(t-\alpha)]\,\mathrm{d}\lambda\,.$$

By comparison with (87) (for a lag filter) we see that all the quantities in the integrand of the extreme right-hand integral in (113) are the conjugaties of the corresponding ones in the integrand of (87). Since in (87) $\psi(t)$ is real; the conjugation of the quantities in the integrand leaves the result of integration unchanged. In other words

$$\psi(t) = \int\limits_{-\infty}^{\infty} \frac{\Phi_{im}^{-}(\lambda)}{\Phi_{ii}^{+}(\lambda)} \exp\left[-j\lambda(t-\alpha)\right] d\lambda \,. \tag{114}$$

Therefore (113) is reduced to

$$\varepsilon_{\min} = \varphi_{mm}(0) - \frac{1}{2\pi}\int\limits_{0}^{\infty} \psi^2(t)\, dt \,. \tag{115}$$

This is a simple expression which permits the determination of the minimum mean-square error without going through the determination of the optimum system. The most important factor in the expression is the ratio

$$\Phi_{im}(\lambda)/\Phi_{ii}^{-}(\lambda) \,. \tag{116}$$

For convenience we shall let $\psi_0(t)$ be $\psi(t)$ when $\alpha = 0$. In terms of $\psi_0(t)$, (115) is

$$\varepsilon_{\min} = \varphi_{mm}(0) - \frac{1}{2\pi}\int\limits_{-\alpha}^{\infty} \psi_0^2(t)\, dt \,. \tag{117}$$

Since $\psi_0(t)$ is always positive we conclude from this expression that the minimum mean-square error decreases with increasing lag. In other words, the performance of a lag filter improves with increasing lag. This is a very interesting and important result. As the lag tends to infinity the minimum mean square error tends to an error that cannot be removed by the linear system. We call this error the irremovable error ε_{irr}. Hence the irremovable mean square error is

$$\varepsilon_{\text{irr}} = \varphi_{mm}(0) - \frac{1}{2\pi}\int\limits_{-\infty}^{\infty} \psi_0^2(t)\, dt \,. \tag{118}$$

It is quite easy to see that if we consider a lead filter than the minimum mean square error is given by (117) with $-\alpha$ replaced by $+\alpha$ that is

$$\varepsilon_{\min} = \varphi_{mm}(0) - \frac{1}{2\pi}\int\limits_{+\alpha}^{\infty} \psi_0^2(t)\, dt \,. \tag{119}$$

For pure prediction this expression can be further simplified. From (119) we find that for long time prediction with $+\alpha$ tending to $+\infty$, the minimum mean-square error tends to the message power. This means that the output of the prediction filter tends to zero as the prediction time tends to infinity.

7. – A general method for expressing the desired output.

In some filter problems the desired output may not be the message but rather the derivative of the message. One might want the integral of the message. A general method for expressing desired outputs of this type is to say that the desired output is the message that has been given a linear operation. One example of this is differentiation and another is integration. If we let $G(\omega)$ be the linear operator in the frequency domain and $g(t)$ be its transform then

$$(120)\qquad f_d(t) = \int_{-\infty}^{\infty} g(\sigma)\, f_m(t-\sigma)\, \mathrm{d}\sigma\,.$$

To introduce this desired output into the general expression (80) we need $\varphi_{id}(\tau)$, which is

$$(121)\quad \varphi_{id}(\tau) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} f_i(t)\, f_d(t+\tau)\, \mathrm{d}t = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} f_i(t)\, \mathrm{d}t \int_{-\infty}^{\infty} g(\sigma) f_m(t+\tau-\sigma)\mathrm{d}\sigma =$$

$$= \int_{\infty}^{\infty} g(\sigma)\, \mathrm{d}\sigma \lim_{T\to\infty}\frac{1}{2T}\int_{T}^{T} f_i(t) f_m(t+\tau-\sigma)\, \mathrm{d}t = \int_{\infty}^{\infty} g(\sigma)\varphi_{in}(\tau-\sigma)\, \mathrm{d}\sigma\,.$$

By transformation of both sides of this equation we have

$$(122)\qquad \Phi_{id}(\lambda) = G(\lambda)\, \Phi_{im}(\lambda)\,,$$

and (80) becomes

$$(123)\qquad \varphi(t) = \int_{-\infty}^{\infty} \frac{G(\lambda)\, \Phi_{im}(\lambda)}{\Phi_{ii}^{-}(\lambda)} \exp[j\lambda t]\, \mathrm{d}\lambda\,.$$

For example, if the desired output is the derivative of the message with lead α, then

$$(124)\qquad G(\lambda) = j\lambda \exp[j\alpha\lambda]\,.$$

In specifying the desired output in this manner we are actually demanding the linear filter system to differentiate the message in the presence of noise, and to predict the result, all in one step and with the least mean square error. A method that can handle a problem of this type is indeed a powerful one.

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3° Trimestre

« Learning » Filters, Predictors and Recognizers.

D. GABOR
Imperial College - London

1. – The principle.

The idea of filters optimizing themselves by « learning » and the basic mathematical formalism was first made public in my report on *Communication and Cybernetics* to the International Symposium of Electronics and Television at Milan, April, 12, 1954 (reprinted in *IRE Proc.*, Vol. CT-1, 19, (1954)). The realization was held up by the difficulty of finding collaborators, and also by the lack of suitable analogue multipliers, which form an essential part of the scheme. Recently KUHRT and HARTEL have developed remarkable analogue multipliers based on the Hall effect in the new semiconducting material indium arsenide, (F. KUHRT: *Eigenschaften der Hallgeneratoren*, in *Siemens Zeits.*, **28**, 370 (1954); W. HARTEL: *Anwendung der Hallgeneratoren*, in *Siemens Zeits.*, **28**, 376 (1954)). The time now appears ripe for attacking the problem in earnest.

The principle described in my paper l.c. may be briefly recapitulated. It is based on the fact that in a limited *waveband* of width F the information can be considered as arriving in the form of discrete data; one every $1/2F$ seconds. From this follows the important consequence that *all operators in a finite waveband can be represented in algebraic form*, operating on $2F$ data per second. It is convenient to take as data the samples of the signal amplitude s_n, at discrete sampling points $0, 1, ..., n$, spaced by $1/2F$, where 0 corresponds to the present instant and n is positive in the past. (Amplitude samples are convenient for easy illustration, but one can as well take as data the expansion coefficients of the signal in terms of any set of suitable functions, such as described in GABOR, l. c.).

Let O be the operator of which the filter is a physical realization. It operates on the past samples of the incoming signal $s(t)$, *i.e.* on s_0, s_1, s_n, s_N.

It is convenient, (and of course practically unavoidable), to assume that beyond a certain time $N/2F$ the past does not matter. We then write the *most general* transformation which a filter can effect in the form

$$O(s) = \sum_0^N s_N r_N + \sum_0^N \sum_0^N s_{n_1} s_{n_2} r_{n_1 n_2} + \sum_0^N \sum_0^N \sum_0^N s_{n_1} s_{n_2} s_{n_3} r_{n_1 n_2 n_3} + \dots 1 .$$

The r_n ... are the *response coefficients* of 1st, 2nd... order. They will be assumed as constants. It may be noted though that in the « most intelligent » filter they would have to be made dependent on past information. Such a filter would notice for instance, from information received, possibly further back than N, that the language to be filtered is English or French, and would adjust its response coefficients accordingly. We leave such universal filters out of consideration from what follows, and consider only filters adapted to *one stationary time series*. In this case the r_n ... are constants, and translation invariant, as they must be. Optimization is then a matter of adjusting the set of the r_n.

The rule for optimization depends on the problem, and on the « success criterion » which one wishes to adopt. There are 3 types of problems.

1) *Filtering*, *i.e.* removing the unwanted component (noise), from the signal received.

2) *Prediction*, *i.e.* producing from the past values $s_0 \dots s_N$ a value with a negative index, say s_{-d}, which is most nearly right in most cases.

3) *Recognition*, *i.e.* producing a distinctive output for a certain type, or types of signals. In the simplest case of *one* type of signal only, this is the same all filtering; all unwanted signals are rejected as noise. The problem becomes distinct if there are several types of signal to be distinguished. The discussion of this problem will be left for later.

There are also various success criteria which one could adopt. One could for instance try to maximise the rate of information, as defined by SHANNON. This has the disadvantage that one cannot even formulate it as a rule for determining the filter parameters before one has solved the recognition problem, because Shannon's definition is in terms of communication *signs* and their probabilities. Moreover, even in the simplest cases, in which one can give a simple mathematical description of the communication signs, the optimization becomes hopelessly complicated. We can therefore consider only criteria which are simple functions of the true signal amplitude $s(0)$ or, in the case of prediction, of $s(d)$ and of the value $O(s)$ supplied by the filter. One could think *e.g.* of the integrated absolute deviation of the value $O(s)$ from the true value. But all these criteria, with the exception of one, are not only clumsy to handle mathematically, but one can never be sure that there will be a *unique* optimum.

The one exception, the one which which we adopt is the *minimum mean square error* criterion, as used by WIENER and by KOLMOGOROFF in their classical work on optimum linear filters. This is easy to handle, and is certain to give a unique solution.

Consider for instance the prediction problem. For simplicity we write s_d instead of s_{-d}, for the value to be produced by the filter, so that d is positive if it represents a delay, negative for the case of prediction. We have then for the mean square error

$$(2)\qquad \overline{(O(s)-s_d)^2} = \sum_0^N\sum_0^N \overline{s_{n_1}s_{n_2}}r_{n_1}r_{n_2} + \sum_0^N\sum_0^N\sum_0^N \overline{s_{n_1}s_{n_2}s_{n_3}}r_{n_1}r_{n_2n_3} +$$

$$+ \sum_0^N\sum_0^N\sum_0^N\sum_0^N \overline{s_{n_1}s_{n_2}s_{n_3}s_{n_4}}r_{n_1n_2}r_{n_2n_3} + \ldots - 2\sum_0^N s_d\overline{s_n r_n} - 2\sum_0^N\sum_0^N \overline{s_d s_{n_1}s_{n_2}}r_{n_1n_2} - \ldots + s_d^2 \,.$$

This is a quadratic form, moreover a positive definite quadratic form of the variables r_n ... hence it must have a minimum, and this is unique. By this we are assured that if we use the least mean square criterion in a « learning » *i.e.*, self-optimizing device, it will not hook itself on to a subsidiary minimum.

The coefficients of the binary products of the r_n ... are all of the form

$$\overline{s_{n_1}s_{n_2}\ldots s_{n_k}}\,.$$

These are the *autocorrelation coefficients* of the stationary series s, of different order. We take the order as one less than the number of factors, so that the *ordinary autocorrelation function*

$$\overline{s_{n_1}s_{n_2}} = \varphi_1(n_2-n_1)$$

will now be called « of the first order ». The autocorrelation functions are all translation invariant, *i.e.*, they are functions of the differences in the n_j only. Hence the k-th order autocorrelation function is

$$(3)\qquad \overline{s_{n_0}s_{n_1}\ldots s_{n_k}} = \varphi_k(n_1-n_0,\; n_2-n_1 \ldots n_k-n_{k-1})\,.$$

In the filtering problem the cross-correlation functions between the signal and the noise will also appear if the noise is additive (see GABOR, l. c.). In either case, carrying out the minimization of the square error one is led to a set of linear equations, with the correlation coefficients, (*i.e.*, values of the correlation functions), as coefficients, and the r_n as the unknowns.

Solving these equations, (say a hundred), is in itself a formidable proposition, but collecting the coefficients, whose number is of the order of the square of the number of equations is practically hopeless. I have therefore proposed, in 1954 a short cut through these difficulties, by suggesting a machine which adjusts itself optimally by a learning or training process. This is much the same as what the higher animals are doing. The brain is, among other things, an astoundingly universal filter. The lower animals are born with « built-in » reflexes, ready for all everyday emergencies, but the immensely more complicated humain brain contains at the start little more than potentialities. It is, as it were, an enormously complex network, in which almost all the switches are still open. This may well be due to the simple fact that the gametes are unable to carry the information for the detailed connections, but however this may be, the fact that we must learn almost the whole analysis of the sensory data and the reactions to almost all but the most elementary situations pays a rich dividend in adaptability. It appears that our machines have now also reached the degree of complication at which it does not pay any longer to plan all their reactions in advance. Instead we must make them potentially capable to cope with every situation, but leave it to them to acquire their reactions by « learning ».

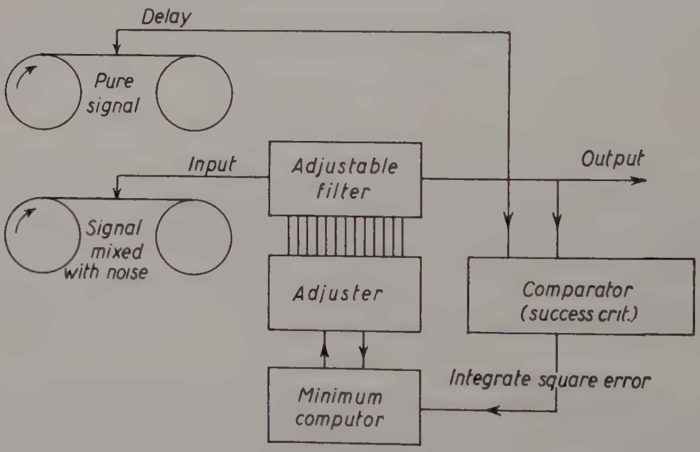


Fig. 1.

Fig. 1 illustrates the « training machine » and is takenf rom the original paper. The filter has a great number of knobs, each representing one of the coefficients r_n ... adjustable *e.g.* by means of a potentiometer. There are two records provided, one of the pure signal, the other of the signal mixed with noise. (In the case of prediction training *one* record will do, with two pickups). The mixed signal is fed into the filter and the output compared with the pure signal, (or « fair copy ») by a « comparator » which calculates the square of the difference, and integrates it over the whole playing time. The adjustment mechanism comes into action only when the whole record is played through.

Various strategies can be adopted for the adjustment of the great number of knobs, say 100. The simplest, though not necessarily the most efficient strategy is Southwell's *relaxation routine*: only one knob is turned at a time. Starting from some first guess, one knob is given successively three positions, say positive maximum—zero—negative maximum. This knob is reset to the original « first guess » position, and the same play repeated in turn with all the other knobs. The minimum computer now calculates the « influence curves » of every knob. In our case these will be simple *parabolas*, which are determined by 3 points, hence as the original setting is one point, two other settings are sufficient. The minimum computer gives the best setting for every knob (if this alone is turned of course, the others remaining in the first guess position,) and indicates the one whose optimum setting gives the *largest* reduction in the integrated square error. In a fully automatic machine it will also turn the knob. The play is now repeated, until no worth-while further reduction can be obtained.

This is a safe but slow routine, it requires 200 runs for one adjustment if there are 100 knobs. There is no doubt that it can be greatly speeded up. First, one will certainly notice after the first 200 runs that only a small fraction of the coefficients is effective. One can then discard all these which give, may be, less than 5% of the largest reduction, and return to them only when all the more important adjustments are made. Second, after the first 200 runs one can assume as a working hypothesis that the order of importance of the others has not changed by setting the most important coefficient. One therefore goes through these in their original order of importance, setting them, after 2 runs in each case, to their relative optima. This procedure is also perfectly safe, because in the case of a positive quadratic form one is certain to proceed towards the absolute minimum so long as every step produces a decrease. Thus, in this method, it takes at most 400 runs before all 100 knobs are reset. One than starts the routine again.

It will be of interest to investigate whether the alternative strategy of « steepest descent », proposed by G. Temple in 1938, may not prove superior to Southwell's relaxation routine. In this method, after taking a small section of the influence curves which can be considered as straight, a computation is made of the linear combination of alterations which produce the largest decrease.

I propose to take the 2 (or for safety perhaps 3) points which determine an influence curve not in 2 (or 3) successive runs, but in one, by doubling (or trebling) the comparator. This means 100 runs for the first re-setting of 100 knobs, and I surmise that it will be seldom necessary to repeat this more than 10 times, making a maximum of 1000 runs in all. As will be shown later, the machine can be made fast enough to process at least 100 data per second, 10000 data in a record of 1 min 40 s duration. 1000 runs can be made therefore in somewhat less than 24 hours (or a little more, taking into account the re-winding time). This is certainly not prohibitive if the machine is fully automatic.

2. – Fields of application.

The learning filter can be considered either as a scientific instrument for solving mathematical problems, or as a prototype model of practical filters, which are functionally (not necessarily structurally), copies of it. These copies do not learn; the prototype has done all the learning for them.

Applications of the second type are evidently the more important ones from a commercial point of view, but it will be better to suspend opinion for a while on the commercial value of this whole work. It is quite conceivable for instance that one will find that non-linear filters are not sufficiently efficient in telephony for justifying the probably very considerable costs. On the other hand, there can be no doubt about the scientific interest of the venture. The learning non-linear filter can be expected to make a long dash into a field of which only the outer fringes have been explored by mathematicians. It may well be that the result of the exploration will be that nothing much less complicated than the human brain can be of appreciable value for assisting the humain brain and senses. But even if the report should be « nothing but desert land for the 100 next miles », at least it will save expeditions equipped rof a score of miles only.

I will classify the possible fields of application into: Communications, Control problems and Statistical forecasts.

3. – Communications.

1) *Telegraphy.* The recognition of Morse signals in a noisy background is a promising field for non-linear filters, because the constant level of the signals is a recognition index which is missed by linear devices.

2) *Radar.* It will be most interesting to investigate whether a system exists better than « correlation reception ». In this system the received noisy signal is multiplied by the emitted pulse, with different delays, and there is a peak when this delay is equal to the return time of the pulse.

The learning filter is very well equipped for dealing with this problems, because correlation reception can be achieved by using its linear terms only. Let the radar pulse correspond to the sequence $x_0 \dots x_n \dots x_M$. We then make

$$r_n = kx_n$$

and the filtered signal is

$$k_0^N x_n^2 ,$$

which represents correlation reception if $N = M$ (this is several thousands in radar, but a smaller number will do in model experiments). Thus the learning filter can realise correlation reception by using its linear terms only, and it would be surprising if its higher order terms could not improve on it.

3) *Telephony.* Here we must distinguish several problems of increasing complexity.

3a) *Filters for eliminating noise from speech.* Human speech has a certain recognizable character for the ear; one can recognize it as having issued from a human mouth long before one can understand it. (Certain gurgling and clicking African languages excepted!) The best that linear filters can do is to use filters of the average spectral characteristic of the language. (Very carefully studied by the G.P.O. who have taken the spectra of almost all European languages.) But this is certainly not all one can do. The difficulty is only that these characteristics are almost certainly in the syllabic modulation, in the slow rythm of amplitude and frequencies, not in the wave-shapes or even in the short-time spectra. I think that little would be gained by trying out the filter simply on speech records, mixed with noise.

It is very likely therefore that in order to be successful, the filter (or a device associated with it), will first have to take the signal to pieces, *i.e.*, analyse speech into the components which have been recognized as essential in the research of the last 30 years, apply the processing to these, and then put the pieces together again.

It is easy to see though, that on the basis of present-day knowledge this would require a very complicated apparatus. The best waveband-saving apparatus at this moment is the Bell 32-channel Vocoder, which separates the spectrum of about 3600 kHz into 32 channels, in each of which a frequency band of 25 Hz is sufficient for transmission. This gives fully acceptable speech, of commercial quality. There are now claims that the 16 channel Vocoder has also reached this level. But even if we accept this claim, and put down the channel frequencies to 20, this still means $2 \times 20 \times 16 = 640$ data per second. One requires at least $\frac{1}{8}$ of a second to recognize the syllabic regularities, which enable the human ear to separate speech from noise, and this still means that the filter has to take in 80 data simultaneously. This is well beyond what one can contemplate for a start.

On the other hand one could well think of eliminating noise with simple characteristics, such as Morse signals and clicks from speech. This must be left to experiments.

3*b*) *Speech telegraphy.* If we drop the criterion of « commercial quality » the problem becomes more manageable, though still difficult. Walter Lawrence has shown convincingly, that perfectly intelligible and human-sounding speech can be synthesized out of 6 slowly changing parameters. (The frequency and intensity of the larynx tone, the intensity of the « hiss », and the position of the first three formants.) Each can be transmitted in a 20 Hz channel, hence the number of data per second is $2 \times 20 \times 6 = 240$. This still gives 30 data every $\frac{1}{8}$ s, but for *speech recognition* it may be sufficient to take 3 consecutive sets of 6 data, 18 in all, which is manageable. These extend over a time of $3 \times 25 = 75$ ms, perhaps one can stretch it to make it 0.1s. This is probably sufficient for recognizing most if not all *morphemes*, such as vowels, which are recognizable in themselves, certain consonants like « sh » which are also recognizable and combinations such as « ka » and « at ».

I have mentioned before that in the simplest case, if only a single type of signal must be recognized, the process is not essentially different from filtering. If for instance we want to recognize « ka », the training record contains the speech record (in some form,) while the « fair copy » or pure signal record has everything wiped out except the « ka »-s which occur in the record. All other speech sounds are considered as noise, and the machine will do its best to eliminate them.

It is easy to see that no *linear* filter could achieve this for speech recognition. In the linear case the problem is this: Given a wanted signal with the components $x_1 \dots x_N$ and unwanted signals $y_1^i \dots y_N^i$ find such a set of coefficients r_n. that

$$\sum_1^N r_n x_n \neq 0 \, ,$$

while for all i

$$\sum_1^N r_n y_n^i = 0 \, .$$

This means *orthogonalizing* the unwanted signals to the wanted signal. It gives a set of linear equations, one more than the number of unwanted signals, and one can solve these if this is less than N. But in the case of speech, analysed in Lawrence's way, in 3 consecutive intervals we have only 18 parameters, sufficient to eliminate 17 unwanted signals only. (One can visualize this by imagining each sound-section as a vector in an 18-dimensional space. There are only 18 orthogonal vectors in such a space.)

A non-linear filter however can deal with this situation, if it has enough terms, as it is not restricted to linear combinations. In general higher algebraic forms have the drawback that if the terms balance at one level, they will not balance at all levels. One can avoid this by choosing for recognition only terms of one order, in which any one signal component occurs always with the same power only. For instance one could take bilinear terms only, specifying that for all i

$$\sum_0^N \sum_0^N r_{n_1 n_2} x_{n_1} y^i_{n_2} = 0 \,,$$

while

$$\sum_0^N \sum_0^N r_{n_1 n_2} x_{n_i} x_{n_2} \neq 0 \,.$$

This gives $\frac{1}{2}N(N+1)$ available parameters, a much greater number. For $N = 18$ this is 171, which again is rather too large, but can probably be reduced. Hence it is not hopeless to train a non-linear filter for speech recognition, by the method of filtering out one morpheme at a time. This, however, would give one (rather complicated) filter for each morpheme, a rather impractical proposition.

An improvement can be obtained as follows.

Assume that we have $64 = 2^6$ morphemes to distinguish. We divide up the filter into 6 parts, with 6 outputs, and the training « fair copy » tape has also 6 tracks. These have only three grades of intensity; positive, negative and zero, and contain the morphemes in binary code, zero being left for uncertainty, or unwanted noise. Thus the 6 filters are trained to give one digit each of the code; they give sufficiently strong positive or negative signals if a recognizable morpheme comes along, and zero for uncertainty, noise, or gaps in the speech.

It is by no means clear whether training *must* lead to success, even with a very complicated filter, if the binary code is arbitrarily chosen, because this presupposes that each morpheme has 6 independent binary characteristics. (JACOBSON has shown, rather convincingly, that there are 7, but some may be redundant.) This is a matter for experiment.

I have dealt with speech recognition at some length, but I do not want to conceal my personal opinion that of all robots the « automatic typist » is the most unnecessary, next to the mechanical translator. It is however such a challengingly difficult problem, that it is not easy to avoid getting intoxicated with it.

4. – Control problems.

I will not attempt a systematic discussion, but will pick out three important problems. In the first two of these we want to achieve *prediction*. (It is easy to modify these into noise-filtering problems.)

1) *Straight follower with prediction.* We observe the instantaneous value $s(0)$ of a quantity, *e.g.* the position of a target. This is taken as the input of the filter, while the output goes into a « system » which contains all the controls

and the whole chain of processes which lead to the output quantity. (For instance if the output quantity is the position of a projectile at the later time d when it reaches the range of the target the « system » will include the gun-laying system and the ballistic process.)

Let again be $O(s)$ the operator of the filter and S the system operator, we prescribe

$$SO(s(0)) = s(d) \,.$$

Let S^{-1} be the inverse system operator, so that

$$SS^{-1} = S^{-1}S = 1 \,.$$

We obtain for the filter operator the rule

$$O(s(0)) = S^{-1}s(d) \,.$$

In this form the problem is ready for the training process, if we know the operators S^{-1}. (For instance the input at time 0 of the system necessary to achieve a certain result at the later time d.) The difficulty might arise that we know S, but we do not know S^{-1}. In this case the filter itself can be used to solve the problem. We must only give it the instruction

$$SO(s) = s$$

and the solution is

$$O = S^{-1} \,.$$

But if we have an analogue of S, we can also put it in series with the filter and solve directly the first equation

$$S(O\,s(0)) = s(d) \,.$$

2) *Follower with feedback.*

We now have the equation

$$S(O\,s(d) - s(0)) = s(d)$$

or

$$s(0) = (O + S^{-1})\,s(d) \,.$$

$O + S^{-1}$ is the inverse of the prediction operator; it produces $s(0)$ from the later value $s(d)$. There is no need to solve it. If we possess an analogue of the system we realize the connections as shown in the above sketch and have again a straightforward prediction-training problem.

Both problems can be at once converted into filtering problems without prediction if we replace $s(0)$ by signal + noise, and $s(d)$ by the « pure signal », which in this case means the true value of the quantity of interest, which we want to achieve, irrespective of the misleading observation.

3) *Analogues of systems which are only experimentally known.* This can be called « Tustin's Problem » as its great importance has been pointed out by A. TUSTIN: *The Mechanism of Economic Systems* (Appendix to 2nd ed. (London, 1955)), and who has also inspired the only two investigations on this subject which are known to me:

J. B. RESWICK: *Determining System Characteristics from Normal Operating Records,* in *Control Engineering* (June 1955).

T. P. GOODMAN: *Experimental Determination of System Characteristics from Correlation Measurements,* M.I.T. Thesis, June (1955) this contains also the full literature of the subject.)

As far as I know, these previous attempts relate to linear systems only, and the imitation of system characteristics is not automatic. The problem is a straightforward one for the learning filter. The system input, as given by operating records, is used as input, and the system output as the « fair copy ». The filter than converts itself into an analogue of the system.

Applications to economic systems are particularly interesting. It is only to be hoped that sufficiently extensive records can be obtained. If one wants to test the prediction value of a filter which can exactly reproduce, say, 100 data, one ought to have test records of at least $1000 \div 10000$ data.

5. – Statistical forecasts.

This is mostly covered by the last section. I want only to mention the particular interest in *weather forecasts.*

The general problem of weather forecasting is roughly as follows. One observes n quantities of interest (such as temperature, air pressure, humidity, wind strength and direction, etc.) in N observation stations. In addition one has parameters such as the time of the year, sunspot activity, etc. From these one wants to forecast n quantities in at least one place. This n-vector is therefore a stochastic function of an $n \times N$-vector, and nN is a very large number. I propose to give our machine 18 inputs, so that we could cope at most with $n = N = 4$ or $n = 3$, $N = 6$. It is hardly possible to say in advance whether this gives the machine a reasonable chance, even if we train it on time series in which the parameters (time of the year, sunspots, etc.) are chosen as carefully equal to the present situation as possible. There would be little point in the machine embodying the laws of air dynamics at this stage, as the data are hopelessly inadequate for dynamical determination. I think though that there is a chance that we might get a good idea of « how complicated a machine would be adequate to deal with the problem? ».

A problem more suitable for a relatively small machine is the one of the *degree of correlation* between two statistical series. NORBERT WIENER has in the last years worked out a most important measure for the *linear* correlation of two series. Our machine is in principle capable of tackling this problem, by using only its linear part, but making the r_n terms numerous enough to span the longest delays between an event in the first series and in the second which may be of any importance. In principle it can, however, measure correlations to higher degrees. The measure of success is always the reduction of the mean square error in the reproduction of a series B by knowledge of the series A, *i.e.* taking this as the input. One can than take different series A', A'' ... and thus quantitatively *allocate the causation* of B to A', A'', etc. One can also test whether for instance A was more the cause of B or vice versa. These are the types of problems which NORBERT WIENER has promised to investigate in his as yet unwritten book on *The Grammar of the Semi-Quantitative Sciences.* I think that there is a good chance of tackling them experimentally.

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3° Trimestre

Television Compression by « Contour Interpolation ».

D. GABOR
Imperial College - London

1. – The principle of contour interpolation.

The enormous waveband of $(3 \div 6)$ MHz which is required for acceptable television pictures has been often contrasted with the limited intake of visual perception, estimated by YVES LE GRAND and others to be of the order of a score of bits per second. At present we can hardly dream of realizing this enormous potential compression ratio. It would probably require an analyser with a complexity comparable to that of the human nervous system, breaking down the raw data into familiar patterns (« Gestalten »), and filling in missing details from past experience. Contour interpolation is a modest first step in this direction. It takes in *two* data at a time (to be compared with one datum in ordinary television and probably a few thousand in the case of the eye (*)), and recognizes only one type of pattern; an outline, or more exactly a short bit of an outline which can be considered as straight without serious error.

The progress from one sensing spot to two might not be very important, were it not that contour interpolation steps in at the weakest point of present day television systems, in which they are most wasteful. *First*, one of the two interlaced fields in TV pictures adds very little to the information; it is indispensable only for suppressing the flicker. *Second* the frame frequency (25 frames/s in Europe, 30 in the U.S.) is higher than justified by the time resolution of the eye; it is also justified only by the flicker. Films with 16 frames/s (with a shutter frequency of $48\ \text{s}^{-1}$) are quite satisfactory if the ratio of bright to dark phases is sufficiently large, and in systems with « optical equalization » in which one picture fades continuously into the next, the frame frequency can

(*) Cfr. the remarks of W. A. H. RUSHTON at the *First London Symposium on Information Theory* (1950), on hundreds of rods connected to a single nerve fibre, a fact which suggests that « Gestalt » formation starts in the eye itself.

be reduced to 6 s^{-1} or less. But, as will be shown, contour interpolation represents considerable progress for the display of motions over the best optical equalization systems. Hence it may not be too sanguine to expect from it a saving ratio of $2\times4=8$; 2 from the saving of every second field, 4 from the saving of 3 frames in 4.

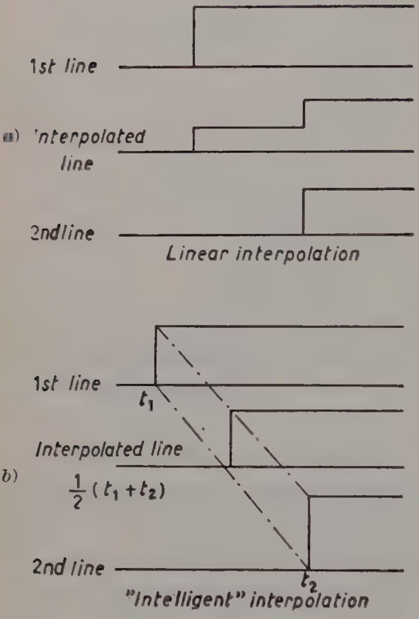


Fig. 1.

Optical equalization is a case of ordinary or linear interpolation, which is illustrated in Fig. 1*a*). The two lines can be equally well considered as next-but-one lines in one frame, or as corresponding lines in next-but-one frames. If the signal amplitudes are $s_1(t)$ and $s_2(t)$ in lines 1 and 2, the interpolated signal is

$$\tfrac{1}{2}s_1(t)+\tfrac{1}{2}s_2(t)\,.$$

It is evident that this preserves clear outlines only if they are vertical and stationary. Slanting contours and moving edges will be blurred. (Nevertheless I think it highly likely that linear interpolation could be used for reducing the frame frequency to one-half of the usual value.)

« Intelligent » or « contour interpolation » is illustrated in Fig. 1*b*). If a step function starts in the first line at t_1 and in the second line at t_2, the interpolated signal is

$$s_{\text{int}}(t-\tfrac{1}{2}(t_1+t_2))=\tfrac{1}{2}s_1(t-t_1)+\tfrac{1}{2}s_2(t-t_2)\,.$$

That is to say the interpolated signal is again a step function which starts at the interpolated time $\frac{1}{2}(t_1+t_2)$ and has the mean amplitude of the steps s_1 and s_2. This is evidently the best guess in the case of outlines which are not too capricious, and it is the correct guess if the outline is that of a uniformly moving body.

The question is only how to find corresponding points (x_1, x_2 or t_1, t_2) in two lines? One might think that it is necessary first to correlate the two lines, and then to make two spots move along with uneven velocities, so that they arrive simultaneously at corresponding points, in which case one has only to take the mean of their simultaneous values. There is, however, a better way, without any intermediate storage of data.

Fig. 2 illustrates this method, which may now be referred to as « contour interpolation » in the case of interpolating between two lines of a frame. The top sketch shows the contour, in reality a narrow zone in which the amplitude changes steeply. It will be a matter for further investigations to decide when a contour is sharp and important enough to set the contour interpolating mechanism into action. (It is certainly worth-while to make it act for the outlines of a head, but not for every hair. If it is triggered off too easily, mistakes are more likely. This is a matter for compromise).

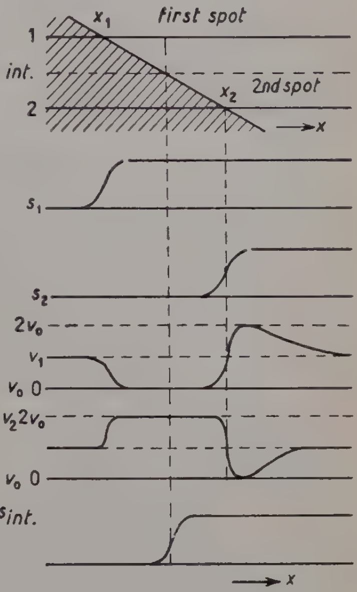


Fig. 2.

Assuming now that the contour is important enough, *e.g.* that the gradient measured over a set distance exceeds a certain absolute value. Up to this point the two scanning spots have proceeded in one vertical line, and at equal speed v_0. When one of them (in Fig. 2 the upper spot), reaches the contour, it stops dead, while the second proceeds at double speed, $2v_0$ until this too reaches the contour. At this moment the second spot stops dead and the first jumps up to double speed. Subsequently the first spot is slowed down gradually, while the second accelerates until the two spots have again aligned themselves vertically, and proceed together with equal speed. Note that during this whole process their mean velocity $\frac{1}{2}(v_1+v_2)$ was constant and equal to v_0.

It is now not sufficient to take the simultaneous mean values of the signal intensities, this would again produce a blurred outline. Instead we must adopt the following rule of *signal distribution* between the two scanning spots: When they proceed together, the amplitudes are added in equal proportions, *i.e.* one half of s_1 is added to one half of s_2. At the instant when the first spot has reached the contour line *the whole signal is transferred to the one which is still moving, i.e.* we take now $s_{\text{int}} = s_2$, and maintain this until the second spot has also caught up with the contour. This procedure will give the correct result if the area at the left of the contour is structureless, or if it has a vertical structure, or, in the case of interpolation between frames, if this area represents a stationary background of any type, and the contour is the outline of a moving object. If these hypotheses fail, for instance if the contour is that of a person moving before a background of flowery wallpaper, the error is not important, because if the structure of the background was not sufficiently pronounced for triggering off the contour mechanism, the eye will not be very critical, and will concentrate on the sharp, moving outline.

As soon as the second spot, moving at double normal speed, has also caught up with the contour, it stops dead and the first spot starts moving with double speed. The amplitude distribution is now reversed, and the rule is simply that *each spot contributes to the interpolated amplitude in proportion to its velocity.* Hence the velocity plots in the diagram on the previous speed represent at the same time transmission or « weighting » functions for the two signal amplitudes.

But what if the second spot misses the contour? In this case, as illustrated in Fig. 3 it is preferable to modify the routine a little. The control mechanism is met in such a way that if on a distance $\Delta x_{\max}$ from the stop of the first spot the second has not found the contour, the two velocities automatically return to their standard values, but *without an overshot.* The effect will be as if the contour had terminated below the first spot, but this is not as bad as if a spurious spot had appeared in the interpolated line at a distance $\frac{1}{2}\Delta x_{\max}$ from the stopping point x_1.

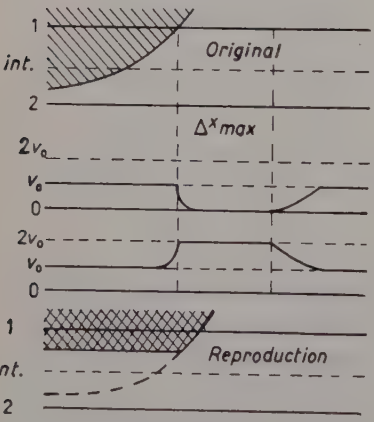


Fig. 3.

It is a matter for experimental research to decide whether this routine (without overshooting), shall not be adopted also in the case when the second spot hits the contour. I see a certain advantage of the overshoot-routine in the point that it gives the correct amplitude in the contour itself, (as at a point in or very near to the contour the signals are divided (50-50) but this may be a minor advantage).

Determining the distance $\Delta x_{\max}$, (or in the case of interpolation between frames the corresponding time interval), is of course an important matter, where a compromise must be struck. In the case of interpolation between lines of a field we ought to make this distance, beyond which the second spot gives up searching for a contour, equal to 6-8 point widths. This means that we interpolate only contours which are inclined more steeply than 1:1 or 1:4 to the horizontal. This is almost certainly sufficient, and the probability that there will be a *spurious* contour in such a narrow interval is not great if we make the critical (triggering) value of the intensity slope large enough.

In the case of interpolation between frames this is a more critical and delicate matter. $\Delta x_{\max}$ now corresponds to the maximum distance by which a contour has moved between two transmitted fields. In other words $\Delta x_{\max}$

sets an upper limit to the horizontal speed of moving objects for which our method is still effective. If we leave out every second frame, the time interval is .08 s, if we transmit one frame in four it is 0.16 s. Television producers tend to avoid fast movements, and a person crossing the screen in two seconds is probably the extreme case we need cater for. Taking the width of the TV screen as equal to 530 picture points, this means that an object of maximum speed moves 265 points in 1 s, 21 points in 0.08 s and 42 points in 0.16 s. It is a matter for research to decide whether the chance of hooking on to the wrong contour will not be too great if we make $\Delta x_{\max} = 20$ or 40 points. Perhaps it will be necessary to call in some further criteria if we want to identify corresponding points at such large distances.

(Ball games will evidently not bear much compression, they take the present standards of 50 fields per second almost to the limit. Moreover tennis games are usually photographed from behind and above, so that the ball rises or fall across 50-100 lines in less than a second, and horizontal interpolation is of course unable to deal with this.)

2. – Functional elements of contour interpolation systems.

It will be useful to list first the functional elements:

1) *Input store.* Data which arrive in succession through the transmission line must be available simultaneously. (*E.g.* two lines in one field, or corresponding lines in two fields.) This necessitates some sort of storage organ.

2) *At least two sensing organs*, such as moving spots, with an organ which decides when the contour mechanism shall be put into action.

3) *Velocity modulating organ*, put into action when the previously mentioned organ decides that a contour has been encountered.

4) *Signal distributor*, which weights the data from the two sensing organs according to the scanning velocities, and adds them.

5) *Output store.* This is necessary unless the interpolated lines, fields or frames are immediately transmitted.

I propose testing the principle first in a rather slow device, suitable for *picture transmission*, at a speed about 1/1000 below that of television, which will be later described briefly. This is of interest in itself, because if it works satisfactorily it could be associated with conventional picture transmitters, such as used by newspapers for news photos, and would enable the users to buy only half of the radio time, perhaps even less. In this mechanical-optical device the input store is a photograph, containing only every second line. The velocity modulator and the signal distributor are mechanical, actuated by photocells. The output store is a film or photographic print.

The chief practical difficulty is in the stores. These will have to be either storage tubes, or magnetic drum storage systems. Storage tubes are available, but very expensive. Magnetic drum storage is very interesting, but will have to be devopled

When starting on a venture like the present, it is good practice to think out the final consequences first, as far as possible, at least in general outlines, and fill in the details later. I will therefore describe first in very general terms the most ambitious system of which I dare think in advance; a television transmission system which gives a compression of 8:1 by contour interpolation alone (*).

3. – A television transmission system with 9:1 compression.

The diagram shows a full cycle of the processing of a TV signal which contains only one field in eight, stretched by some means, (not to be discussed here) so as to fill the whole time but only $\frac{1}{8}$ of the normal waveband. In this diagram a single field scan is represented as an unbroken straight line. The explanation is contained in the sketch at the left. The zones allotted to the single lines are shown side-by-side instead of on top of one another. A stored picture is represented by a shaded area. The second field, which is usually interlaced with the other is shown on top of the first field. (In fact there is no need to interlace these in the storage organs.)

The three received fields are recorded in succession in the three storage areas, which may be storage tubes or magnetic drums, and then the cycle starts afresh. As soon as two lines of a field are stored, *line interpolation* starts as indicated schematically in the bottom left diagram only. This again may be done, by means of storage tubes or drums, but these need not hold more than 3 lines at a time.

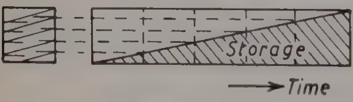


Fig. 4.

There are in every third of the full cycle two full stores, and the between-frames interpolation is carried out between these two, four times in every third of a full cycle. Both fields (the originally recorded and the interpolated one,) are scanned in sequence, so that the result of the interpolation can be immediately transmitted, (without output storage), ready for reception by

(*) This may then be combined with band-compression system by equalization of information rate, such as proposed C. CHERRY and G. G. GOURIET, and F. SCHROTER.

ordinary television receivers. For special purposes, such as cinema or industrial television it may be preferable to transmit the two fields sequentially, and repeat every frame to avoid flicker. This can be done by interlacing the two fields in the storage tube, and scan the whole frame with 405 lines in the

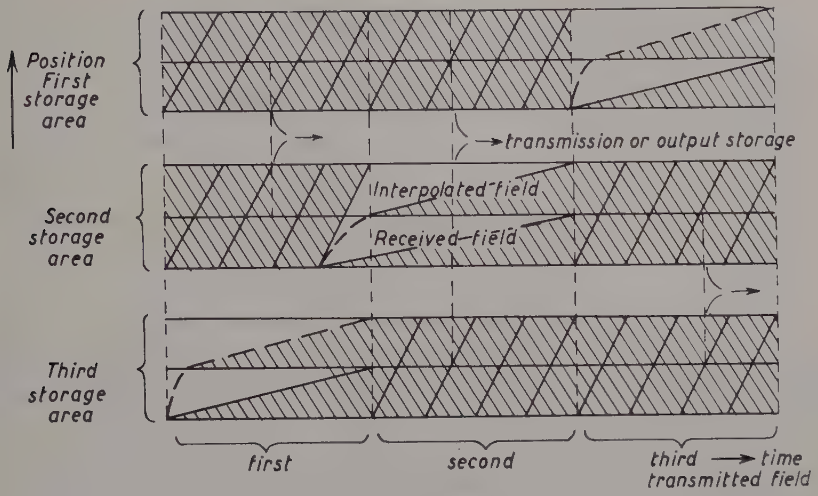


Fig. 5. – 8:1 compression system.

time taken usually for 202.5 lines. (This of course requires twice the waveband, but there is no objection to this in « closed circuit » TV systems.)

The main application of compression systems is of course international in particular transatlantic television, where the cost of terminal equipment is negligible compared with that of the line. Here a compression factor of the order 20 appears desirable. Contour interpolation can contribute to this a factor of 4÷8, and equalization of information rate a further factor of 3÷4, hence the aim does not appear unattainable. Very high perfection is of course expected of every element in such an extensive reprocessing of the information. No noise or position error must be introduced by the repeated storing and reading of the pictures.

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3º Trimestre

Intelligent Behavior in Problem-Solving Machines (*).

H. L. GELERNTER and N. ROCHESTER

International Business Machines Co., Research Center - Yorktown Heights, N.Y.

1. – Introduction.

Modern machines execute giant tasks in arithmetic and carry out clerical operations that are far beyond human capacity, but we have not yet learned to apply them to problems that require more than a barest minimum of ingenuity of resourcefulness. This paper reports some early results in an approach to the problem of learning how to use machines in these presently unmanageable areas. The goal of this research is the design of a machine whose behavior exhibits more of the characteristics of human intelligence.

We shall concern ourselves in particular with a single representative problem; one which contains in relatively pure form the difficulties we must understand and overcome in order to attain our stated goal. The special case we have chosen is the proof of theorems in Euclidian plane geometry in the manner of, let us say, a high school sophomore. It must be emphasized that although plane geometry will yield to a decision algorithm, the proofs offered by the machine will not be of this nature. The methods to be developed will be no less valid for problem solving in systems where no such decision algorithm exists.

Rejecting the application of a decision algorithm as uninteresting (in the case of plane geometry) or impossible (for most problems of interest), there remain two alternative approaches to the proof of theorems in formal systems. The first consists in exhaustively developing the proof from the axioms and hypotheses of the system by systematically applying the rules of transformation until the required proof has been produced (the so-called « British Museum algorithm » of NEWELL and SIMON (**)). There is ample evidence that this pro-

(*) This paper, presented at Varenna, had already been given by the Authors in the June 1958 for publication to the *IBM Journal of Research and Development*, where it effectively appeared in the Vol. II, No. 4, October 1958. (*N.d.R.*).

(**) A. NEWELL, G. S. SHAW and H. A. SIMON: *Empirical Explorations of the Logic Theory Machine*, in *Proceedings of the Western Joint Computer Conferences* (February 1957).

cedure would require an impossibly large number of steps for all but the most trivial theorems of the most trivial formal systems. The last remaining alternative is to have the machine rely upon heuristic methods, as people usually do under similar circumstances.

Problems for which people use heuristic methods seem to have the following characteristic. The work begins routinely, and then suddenly the person experiences a flash of understanding. This is followed by the writing down and checking of the solution. What seems to be happening is that the person first uses heuristic methods to look for a solution. To each suggestion turned up by the heuristic methods he applies some sort of a test. The flash of understanding comes when some suggestion gets a high score on the test. The clerical task that follows is the transformation from « suggestion space » (*) to « problem space ». The transformation is possible, of course, only if a valid solution has been indicated. This is what the geometry machine does.

Instead of geometry we might have chosen a certain class of probability problems, proofs of theorems in projective geometry, proofs of trigonometric identities, proofs in part of number theory, or the evaluation of indefinite integrals. There were, however, compelling reasons for choosing plane geometry, the most important being the readily understood « suggestion space » offered by the diagram (the semantic interpretation of the formal system), and the ease of transforming « proof indications » into problem space. This will be considered in detail later in the paper. An important secondary reason was the fact that everyone who would be interested in our results has studied Euclid, so the results can be communicated more efficiently.

It should be noted here that the geometry project is a consequence of the Dartmouth Summer Research Project on Artificial Intelligence, standing on a foundation laid by the members of the study (**), and evolving from the pioneering work of NEWELL and SIMON in heuristic programming [1].

Not all problems whose solutions seem to be accompanied by a « flash of understanding » are elementary enough to lie within the scope of the methods described below. Many have difficulties of a more profound nature. It will be possible to say a little more about this later, but a secure understanding of the nature of these harder problems will come only after more research has been done.

The explanation of the precise meaning of the term « heuristic method » is an important part of this paper. For the moment, however, we shall consider that a heuristic method (or a heuristic, to use the noun form) is a procedure that may lead us by a quick shortcut to the goal we seek or it may lead us down a blind alley. It is impossible to tell which until the heuristic

(*) A. NEWEL and H. A. SIMON have used the term « planning space ».

(**) Particularly J. MCCARTHY, M. L. MINSKY, and one of the authors (N.R.).

has been applied and the results checked by some formal process of reasoning. If a method does not have the characteristic that it may lead us astray, we would not call it a heuristic, but rather an algorithm (*). The reason for using heuristic instead of algorithms is that they may lead us more quickly to our goal and they allow us to venture by machine into areas where there are no algorithm (**).

One final remark needs to be made. Since people seem to use heuristic reasoning in nearly every intelligent act, it is reasonable to ask why some task more familiar and natural for people was not chosen as representative of the class rather than plane geometry. Several alternatives to geometry were, in fact, considered and rejected for failing to satisfy one or more of the following requirements:

1) The task must include a kind of reasoning that we are not yet able to get our machines to do but about which we have ideas and think we can learn to manage.

2) It must not contain harder kinds of reasoning that are too far beyond our understanding.

3) It must not be cluttered with too much irrelevant work.

Most human acts fail to meet requirement 2). We have a long way to go before our machines can play Turing's « Imitation Game » and win [2].

2. – Geometry.

A standard dictionary defines « geometry » as « the theory of space and of figures in space », and indeed, most people would offer a similar definition. To the mathematician, however, geometry represents a formal mathematical

(*) A decision procedure applied under the constraint of a time limit behaves as if it were a heuristic.

(**) There are classes of problems, for example, proofs of theorems in number theory, where it can be shown that no decision procedure can be devised. Heuristic procedures should enable us to get machines to solve problems that are members of such classes. It should be evident that no set of heuristics together with the programs to employ them can guarantee that a machine will solve every member of such a class. All that a machine can do is to probe around and perhaps come up with an answer. This, of course, is all that people can do. It should be evident, too, that a program utilizing heuristics can perfectly well be an algorithm that is guaranteed to solve any member of some class of problems. Such a class must, of course, be amenable to a decision procedure. The contribution of an individual heuristic here is that it may lead to a short cut. The geometry theorem machine will probably be an algorithm of this kind.

system within which proofs are possible, and which can be related to real space if this seems interesting for the purpose at hand, but which can alternatively be related to concepts having no physical reality or significance. The machine considers geometry primarily as a formal system but uses the interpretation in terms of figures in space for heuristic purposes.

A formal system such as geometry comprises:

1) Primitive symbols.
2) Rules of formation.
3) Well-formed formulas
4) Axioms.
5) Rules of inference.
6) Theorems.

The set of primitive symbols (or alphabet) for geometry are those characters which are interpreted as the names of points together with those interpreted as specifying relations between discrete sets of points, or between a given set and the universe of points (*e.g.*, $=$, $\|$, A, B, Δ). In order to make proofs in geometry it is for example, not necessary to think of a line as something long, thin, and straight. It is sufficient to be able to recognize the symbol « line ».

The rules of formation specify how to assemble the primitive symbols into well formed formulas (statements) which may be valid or invalid within the formal system. For examples, « Two sides of every triangle are parallel » is a well formed formula (although not valid), whereas « Two exists of obtuse every one point » is not a well formed formula. We can ask the machine whether the first is true (interpreting formal validity as truth), but the second is gibberish because it does not obey the rules of formation. These rules are, in a sense, the grammar of a language whose vocabulary comprises the alphabet of primitive symbols.

The axioms are a set of well formed formulas such as « Through every pair of points there can be drawn one and only one straight line » which are selected to serve as a foundation on which to build. They are regarded as being true by definition, if you like.

The rules of inference are the means by which the validity of one well-formed formula can be derived from others that are already established. The new formula is said to be immediately inferred from the given one or set by the specified rule of inference.

A proof is a succession of well-formed formulas in which each formula (or line of proof) either follows by one of the rules of inference from the preceeding formulas, or is an axiom or previously established theorem. A theorem is the last line in a proof.

To recapitulate, a problem presented to our machine is a statement in a formal logistic system, and the solution to that problem will be a sequence of statements each of which is a string of symbols in the alphabet of that system. The last statement of the sequence will be the problem itself, the first will always be an axiom or previously established theorem of the system (*). Every other formula will be immediately inferable from some set preceding it, or will itself be an axiom or previously established theorem.

This simple and elegant description of geometry is essentially the one given to the high school sophomore. It will shortly be seen that this view is too naive to describe what really happens, but for the moment it will be expedient to continue the exposition as if it were true, because the idealization has a significance of its own. There are a number of things to be pointed out about this ideal view of geometry. For one thing, there is a difference between finding a proof and checking it. To check a proof one merely follows some simple rules that are set down very precisely. To discover a proof, on the other hand, requires ingenuity and imagination. One must use good intuitive judgement in selecting which of many possible alternatives is a step in the right direction. The high school sophomore does not have a complete set of explicit rules to guide him in finding a proof.

Since the checking of a proof is a clerical procedure there is no reason why a machine cannot easily do it. A well-formed formula (*i.e.* axiom, line of a proof, or theorem) would be a string of data words in memory, and a rule of formation or of inference would be a subprogram. There is nothing really new or difficult about this, and many programs have been written to make machines do jobs as difficult. The artificial geometer will have a subprogram which is an algorithm for checking a proof.

The process of discovering a proof is another matter, and the question of how to get a machine to do it is the subject of this paper. The student or the machine can be given some useful hints, but must also be provided with a warning that these hints may be misleading. For example, it can be said that if the proposition to be proved involves parallel lines and equality of angles there is a good chance that it will help to try the tehorem:

« If two parallel lines are intersected by a third line, the opposite interior angles are equal ».

This advice is a heuristic that can be given to the machine or student. It will lead to a proof in a good many cases, but will as often lead nowhere at all.

(*) In the case of a theorem contingent upon a set of hypotheses, the proof is developed in an extended system in which the hypotheses are appended to the original set of axioms. The transformation of this categorical proof to the desired hypothetical one is trivial.

Thus far, there has been no mention of drawing figures. It is of course quite possible to discover a proof in a formal system without interpreting that system, and in the case of geometry, except for the need to discover proofs efficiently, or for applying theorems to practical problems, one need never make a drawing. The creative mathematician, however, generally finds his most valuable insights into a problem by considering a model of the formal system in which the problem is couched. In the case of plane geometry, the model is a diagram, a semantic interpretation of the formal system in which, to quote Euclid, the symbol « point » stands for « that which has no parts », a « line » is « breadthless length », and so on. The model is so useful an aid for discovering proofs in geometry that few people would attempt a proof without first drawing a diagram, if not physically, then in view of the mind's eye. If a calculated effort is made to avoid spurious coincidences, then one is usually safe in generalizing any statement in the formal system that correctly describes the diagram, with the notable exception of those statements concerning inequalities.

We cannot emphasize too strongly the following point. To serve as a heuristic device in problem solving, it is not necessary that the model lie in rigorous one-to-one correspondence with the abstract system. *It is only necessary that they correspond in a sufficient number of ways to be useful.* The success of the model in designating correct solutions to problems in that system (solutions that will be checked within the framework of the abstract system) is the only criterion one need apply in judging the suitability of a given model (*).

If the model is indeed a semantic interpretation of a formal logistic system, then it is most desirable that the interpretation satisfy every axiom of the formal system. But should the interpretation be valid too for some richer formal system (or poorer one, for that matter), its heuristic value might be impaired, but by no means eliminated.

3. – Heuristic methods.

The proof of theorems in Euclidian plane geometry in the sense described above requires the extensive use of heuristic methods, and it is these methods rather than geometry that are of primary interest to us. The role of geo-

(*) A. Newell and H. S. Simon, in private communication with the authors, have described an abstract model for a propositional calculus which is not a semantic interpretation but which, in fact, is another formal system in which it is trivially easy to prove the transformed theorems. Since this is a true heuristic, it is not always possible to transform it back to the problem space.

metry is to provide a problem of the right difficulty to permit a thorough development and understanding of the class of heuristic involved.

The steps in a typical application of a heuristic method to theorem proving are the following:

1) Calculate the character (*) of the theorem.

2) Using the theorem character, calculate the applicable methods and estimate the merit of each.

3) Select the most appropriate method.

4) Try it.

5) In case of failure, cross off this method and return to step 3).

6) In case of success, print the proof and stop.

The character of a theorem (or of any problem) is in essence the machine's description of the theorem (or the problem). In its simplest form, the character may be represented by a vector, each element of which describes a given property of either the syntactic statement of the theorem or its semantic representation. The vector designating the applicable methods and estimated merit of each is a vector function of the character. The figures of merit are, of course, only guesses based initially on the experience of the programmer, and subsequently modified by the machine in the light of its experience.

Defining the term *characteristic* as a given element of the character vector, the following might be introduced as syntactic characteristics of a theorem:

C_i = 1 if the hypotheses contain the symbol $\|$, 0 otherwise;

C_j = 1 if the consequents of the theorem contain the symbol $\|$, 0 otherwise;

C_k = 1 if there exists a permutation of the names of points in the hypotheses that leaves the set of hypotheses unchanged, 0 otherwise; and so on.

Examples of semantic characteristics are the following:

$C_l = n$; where n is the number of axes of symmetry in the diagram;

C_m = 1 if two angles of segments are to be proved equal, and they are corresponding elements of congruent triangles, 0 otherwise; and so on.

(*) The term character was introduced by MINSKY (M. L. MINSKY, *Heuristic Aspects of the Artificial Intelligence Problem*, Lincoln Laboratory Report 34-5 December 1956), and is to be understood in its dictionary sense. The particular machine representation of a theorem character selected by the authors differs somewhat from that of MINSKY, but this important concept is due to him.

The rules formalized into the vector function that transforms the character of a problem into a sequence of designated methods of approach and the estimated merit of each will in general fall into two categories. The first will contain those heuristics which operate on the syntactic characteristics of the problem. The second will, in the general case of a problem-solving machine, comprise those rules which operate on the characteristics of the model. For the artificial geometer, these are the semantic characteristics of the model as described above.

The problem of strategy and tactics in choosing methods is most important. One obvious strategy mentioned earlier is to explore all alternatives systematically, and this is known to be inadequate for many problems and is considered by the authors to be uninteresting and probably useless for geometry. The strategy and tactics used by NEWELL and SIMON in their achievement in theorem proving by machines are not adequate for this harder problem on present day machines. Their proofs were at most, three or four steps long and machine time required is probably an exponential function of the number of steps. Clearly the ten step proofs of geometry will require much more selective heuristics than those adequate for propositional calculus.

The authors have at present a system of strategy and tactics. It does not seem useful to report it in detail at this time because machine experience will probably induce major revisions and improvements. It is clear, however, that the skill with which the machine selects and manipulates methods will distinguish a good machine from a poor one. Since it is impossible to predict the detailed behavior of so complex an information processing system as the artificial geometer, it is necessary to write the program and run the simulation before conclusions can be reached with confidence.

The speed with which a difficult problem can be solved is an essential factor in determining the usefulness of an intelligent machine. This speed cannot be achieved by little steps like inventing faster components. On the scale considered here a factor of ten is a minor change in speed. Suppose, for example, that a given proof requires ten steps. If for each step, the machine must explore three alternatives, there will be about 20 000 things to consider. A slightly less intelligent machine that must explore six alternatives will have to consider 20 000 000 things. For problems having longer solutions, selectiveness becomes more important exponentially.

4. – Syntactic symmetry.

The formal system of plane geometry will be a difficult one for the machine to manipulate. Not only are the alphabet and axiom set both large, but geometry must be formalized in the lower functional calculus, at the very least.

The difficulty is compounded, too, by the fact that the predicates of plane geometry exhibit a high degree of symmetry, and a given statement in the system will in general admit a multiplicity of completely equivalent forms.

These symmetries are at times a painful thing to contend with; they make it necessary that a theorem be considered in every one of its equivalent forms in seeking to establish a deduction by means of substitution. On the other hand, they are the basis of a powerful new rule, completely syntactic in nature, that simplifies immensely the search for a proof of a theorem displaying these symmetries. The rule will prevent the machine from searching in a circle for useful intermediate steps, or subgoals, to bridge the gap between antecedent and consequent of the theorem to be proved. In effect, it removes from consideration those subgoals which are formally equivalent to some subgoal that has already been incorporated into the structure of the search for a proof.

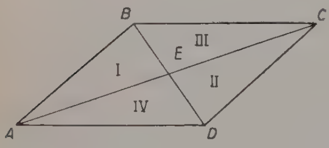


Fig. 1.

We shall introduce the rule by an example. Let us consider the following theorem: The diagonals of a parallelogram bisect one another (Fig. 1).

To solve the problem, the machine must establish the formulas:

$$AE = EC\,,$$

and

$$BE = ED\,.$$

Now it would be most useful if the artificial geometer could recognize, as people usually do, that the proof of the second formula is essentially the same as that for the first, and therefore only one of the two need be established. But it is even more important that the machine fall not into the class of trap illustrated by the following redundant search process. The method chosen is that of congruent triangles, and in order to establish the formula $\Delta\mathrm{I} \cong \Delta\mathrm{II}$ from which the theorem may be immediately inferred, the machine sets at some later stage the subgoal $\Delta\mathrm{III} \cong \Delta\mathrm{IV}$. The geometer will, in fact, satisfy our requirements on both these points. The mechanism whereby this is accomplished is an embodiment of the theorem and rule specified below.

Consider first the following definition: Let π be a permutation on the names of the syntactic variables in a theorem. Then π is a *syntactic symmetry* of the theorem if its operation on the set of hypotheses leaves the set unchanged except for a possible transformation into an equivalent form with respect to the symmetries of the predicates (*i.e.* $\pi\{H\} \equiv \{H\}$ is valid). We can now state the required theorem thus:

If Γ is a well-formed formula provable from the set of hypothese $\{H\}$, then $\pi\Gamma$ is a well-formed formula provable from the same set $\{H\}$. The formula $\pi\Gamma$ will be called the *syntactic conjugate* of Γ. The proof of the theorem is quite trivial, and follows from the fact that the syntactic variables in a theorem may be renamed without destroying the validity of the theorem.

Thus, if $\{H\} \supset \Gamma$ is valid,

then $\pi\{H\} \supset \pi\Gamma$ follows by the rule of substitution.

Since $\pi\{H\} \equiv \{H\}$

$\{H\} \supset \pi\Gamma$ is valid.

The theorem itself grants the machine the same power the human mathematician has at his disposal when he recognizes the equivalence of two different statements with respect to a given formal system, for now it may establish the syntactic conjugate of any valid formula Γ by merely asserting « similarly $\pi\Gamma$ ». The rule of syntactic symmetry follows from the theorem. It is used by the machine to construct, given the heuristics and methods at its disposal, the optimum *problem-solving graph*, and a description of such a graph is in order at this point.

Let G_0 be the formal statement to be established by the proof. It will be called the problem goal. If G_i is a formal statement with the property that G_{i-1} may be immediately inferred from G_i, then G_i is said to be a *subgoal* of order i for the problem. All G_j such that $j < i$ are *higher subgoals* than G_i, where G_0 is considered to be a subgoal of order zero. The *problem solving graph* has as nodes the G_i, with each G_i joined to at least one G_{i-1} by directed link. Each link represents a given transformation from G_i to G_{i-1}. The problem is solved when any G_i can be immediately inferred from the hypotheses and axioms (*).

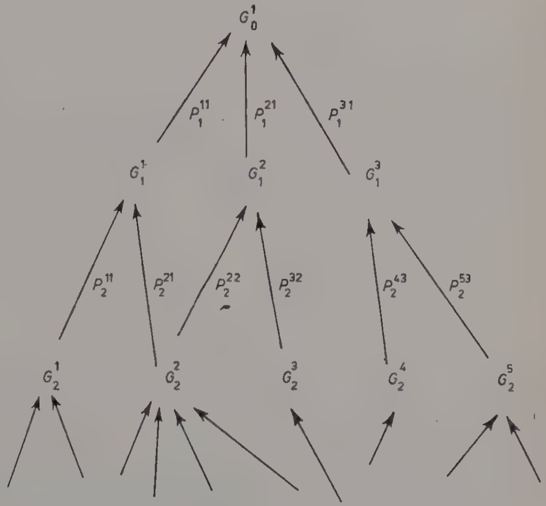


Fig. 2. – The nodes G_i^α represent subgoals of order i, with α numbering the subgoals of a given order; $P_i^{\alpha\beta}$ is a transformation on G_i^α into G_{i-1}^β.

(*) The completed proof will use a deduction metatheorem to get $\vdash \{H\} \supset G_0$ from $\{H\} \vdash G_0$.

We can now specify the rule of syntactic symmetry thus: G_i is not a suitable subgoal to add to the problem solving graph if it is the syntactic conjugate of any G_j for $i \geqslant j$, for any proof sequence leading to G_i is identical with a conjugate sequence leading to G_j with the variables renamed, and any mechanism leading to a proof of G_i would as well prove G_j. If $i = j$, the two subgoals are in effect redundant, and if $i > j$, the sequence leading to G_i leads to G_j when conjugated, and all the steps G_k, $j \geqslant k > j$ can be eliminated.

In the light of the above, we may now re-examine our introductory problem (Fig. 1). The machine must establish the following two goals:

$$G_0^1: \quad AE = EC\,,$$

$$G_0^2: \quad BE = ED\,.$$

By the theorem of syntactic symmetry, the machine will eliminate G_0^2 from the graph, since $G_0^2 = \pi G_0^1$, where π is the transformation A *into* B, B *into* C, C *into* D, *and* D *into* A, and after proving G_0^1, will assert « similarly, G_0^2 ». Then, if at some point in the proof $\triangle ABE \cong \triangle CED$ is a subgoal, it will eliminate the statement $\triangle BCE \cong \triangle DEA$ as a possible subgoal; if $AB = CD$ is a subgoal, $BC = DA$ will be removed from consideration. Clearly every directed path through the problem solving graph from hypotheses to goal will be unique under the π transformation, and will be the shortest one in that it will contain no redundant sub-graphs (no two nodes will be linkable by a π transformation).

Syntactic rules such as the above will be essential to the success of the plane geometry machine. But while they ease the labor of the geometer considerably as it threads a path from problem to solution, they are, except in the simplest cases, powerless to indicate which path, among the very many possible, does indeed lead to a solution, and which wander off into infinity, regressing farther from the goal with each step. The geometer will need more information about most problems before it can even begin to seek a solution. It will find the information as the mathematician does, in the diagram.

5. – Semantic heuristics.

Semantic heuristics are concerned with the body of pertinent and probably true statements that can be obtained by observing the diagram. For example, one of the first such rules to be applied the by geometer in a particular case will be the following:

> If the diagram consists of a « bare » simple polygon, a construction will probably be required.

A rule to indicate which construction to make might be:

> If the figure has one axis of symmetry, and it is not drawn, then draw it.

A most useful rule will be:

> If the theorem asks that two line segments or angles be proved equal, determine by measuring whether these are corresponding parts of apparently congruent triangles. If so, attempt to prove the congruence.
>
> If necessary, draw lines connecting existing points in the diagram in order to create the congruent triangles.

Another frequently used heuristic will be:

> If two apparently parallel lines are crossed by a transversal, attempt to establish the parallelism by considering the angles.

A more complete understanding and appraisal of the appropriate heuristics will be one of the major consequences of experimentation.

It should be clear that the best set of heuristic rules, in other words the set that is the best compromise between conciseness and efficiency, should not be expected to yield the best proof in every case. Indeed, in a number of awkward cases the rules will impede, rather than aid, the search for a concise proof. In some cases the machine will make a construction and produce an elaborate proof while missing a simple elegant proof. People, too, do this. But these awkward cases should be the exception, and the heuristic rules look powerful enough to make an efficient machine.

6. – Rigor.

Mathematical rigor becomes a significant matter in two different aspects of the artificial geometer. One of these is that machines can provide, in a sense, more rigorous proofs than have hitherto been available. More important than this is the second aspect, and this is that the machine is like a good human mathematician, in that it increases its output and improves its communication with other mathematicians by taking chances with rigor.

Axioms and theorems are objects that can be examined and manipulated by people and machines. These present no problem. However, methods of inference are instructions to do something. In the case of machines they are programs of instructions in machine language. In the case of people they are instructions expressed in a natural language and intended to control human behavior. Except for undetected blunders in design of a machine or in the writing of a set of machine instructions, the machine and its instructions are fully understood. And when one of these blunders is detected, it causes merely annoyance and not bewilderment. Therefore when a machine proves a theorem, there is in principle, no doubt about what is going on, and except for possible apprehensions about human blunders or undetected machine malfunctions there is no doubt about rigor.

In contrast the human situation is rather poor. While much is known about the human brain, the basic principles of operation are still unknown. Perhaps some of today's conjectures are correct, but we have no sure way to select the correct conjectures from among the various contradictory proposals. Furthermore, natural languages are not yet perfectly understood and again there are contradictory theories. It therefore seems unwise to rely on the rigor of any system based on such a welter of ignorance.

It is interesting to observe that the most rigorous treatments of the foundations of mathematics seem equivalent to designing a machine and a machine language and henceforth communicating in this language. In one case [3] the mathematician even uses the term « machine », although his machines could not actually be built because they contain parts with infinite dimensions. Other really good treatments do not use the word « machine » but are essentially equivalent. It should be clear then that the translation of a formal system into a machine program is reasonable and natural.

The other aspect of rigor is quite different. Most elementary textbooks on geometry fail to prove betweenness relations. In Fig. 3, the acute angle ABC is bisected by the line segment BD. The line segments BD and AC are extended to infinity, thus becoming lines BD and AC. Point E is defined as the intersection of lines BD and AC. Now how can it be determined whether point E lies between A and C or to the left of A or to the right of C?

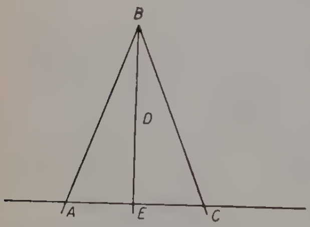


Fig. 3. – Bisected angle.

Ordinarily this decision is made by looking at the figure. In rigorous treatments it is proven formally, but this is a tedious effort and except for metamathematical considerations, not really necessary. Expediency dictates that the mathematician should neglect the possibility that semantic heuristics will lead him astray and get on with the work rather than dally over proofs of betweenness. Because people rarely get in trouble because of honest errors of this kind, traditional geometry excludes proofs of betweenness, and most mathematics appear to lack rigor because many matters are settled by heuristic methods rather than formal proofs. It seems clear that the machine must be able to work this way if it is to become proficient.

The artificial geometer decides questions of betweenness by measurements on the figures. But whenever it does so, it explicitly records the necessary assumptions for a given proof so as to leave a record of its guesses. There is, of course, a danger that the machine will be proving only a special instance of the theorem presented to it, but this danger can be minimized by having the machine draw alternate diagrams to test the generality of its assumptions when they are necessary.

7. – Programming the geometer.

The organization of the program falls naturally into three parts; a « syntax computer » and a « diagram computer » embedded in an executive routine, the « heuristic computer ». The flow of control is indicated in Fig. 4.

The syntax computer contains the formal system and its purpose is to establish the formal proof. Geometry is expressed in a Post-Rosenbloom canonical language which should be useful for a much wider range of formal systems than geometry. The heuristic computer can submit any sequence of lines of proof to the syntax computer which will test them to see if they are correct.

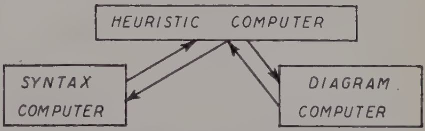


Fig. 4. – Flow chart of the artificial geometer.

The diagram computer makes constructions and measures them. It does this by means of geometry and floating point calculations. However, it keeps all this secret from the heuristic computer and reports only qualitative information of the type acquired by a mathematician in scanning a well-drawn figure. The behavior of the heuristic computer and the syntax computer would not be changed if the diagram computer were replaced by a machine that could draw diagrams on paper and observe them.

The heuristic computer does most of the things that have been discussed in this report. It contains the heuristic rules and decides what to do next. The subordinate computers only follow its instructions and answerits questions.

The program is being written in an information processing language constructed by appending a large set of special functions to the Fortran compiler for the IBM 704. The language increases manyfold the ease of writing programs of the nature of the geometer, and will be reported upon in detail in a subsequent paper.

8. – Learning in intelligent machines.

The machine described thus far will exhibit intelligent behavior, but it will not improve its technique. Except for the annexation of previously proved theorems to its axiom list, its structure is static. A rigorous sequence of practice problems will not improve its performance at all in solving a given problem unless a usable theorem is among them. Such a machine, incapable of developing its own structure will always be limited in the class of problems it can solve by the initial intent of its designer. It seems that the problem of designing a machine of general intelligence will be enormously greater, if at all possible, than designing a not so intelligent one with the capacity to learn.

One might attempt to endow an automaton like the geometer with the ability to learn at various levels of sophistication. Indeed, the behavior of the machine in storing away for future use each theorem it has proved may be interpreted as learning of a rudimentary sort. This might be refined by having the machine become selective in its choice of theorems for permanent storage, rejecting those which do not seem (by some well-defined criteria) to be sufficiently « interesting » or general to be useful later on. Similarly, instead of « forgetting » all lemmas it might have established as intermediate steps in the proof of the theorems offered to it, to be rederived when needed, the machine might select the especially interesting ones for its list of established theorems.

The next level of learning is indicated when the machine adjusts, on the basis of its experience, the probability for success it assigns to a given heuriste rule for a theorem with a given character. This is the learning involved when the machine uses results on one problem to improve its guesses about similar problems. As the geometer is given problems of a given class, say problems about parallelograms, it would get better at handling them. After it had been given a graded sequence of harder and harder problems, its performance should be much better and it could be said to have learned to prove parallelogram theorems. The highest level to which we aspire for an early model geometer will be involved when it looks over the quality of its predictions and discards as irrelevant some of the criteria that comprise the problem character. The earliest models of the geometer will include only low levels; later models will be better.

Beyond these kinds of learning we can see other things. Before we come to them, however, we will probably be working on machines to solve harder problems than those of geometry. There are kinds of learning that are needed only by machines that take their environment more seriously than theorem proving machines do. These will be discussed in the next section. But we can hope that a theorem machine might some day be able to observe that some sequence of methods was effective in certain circumstances, and consequently streamline the sequence into a single method and in this way devise a new method.

But in still another vein there are possibilities for theorem machines. Instead of providing a machine with a formal sytem and a sequence of propositions to prove, it could be given a formal system and be asked to see what it could find. Here it would at least need criteria for the utility of theorems in proving other theorems and for the elegance of a proof in terms of large achievements in a small number of stops. New kinds of learning would be used here.

Before closing the subject of learning machines, there are some further considerations to deal with. A computer is, after all, just a finite automaton, and, as such, its behavior is completely determined by its internal state at the beginning and subsequent input information. This being the case, it can be argued that its response to any set of input signals is in principle predictable and is consequently uninteresting and not worthy of the description « intelligent ». Another version of this objection is the following. The machine, endowed with heuristics and judgements of its designer, is but a trivial extension of that person, in principle no different from a slide rule in the hands of an engineer.

From a certain irrelevant point of view the objection is justified, but in practice the behavior of the machine is far from being predictable. That this is indeed the case is well illustrated by the fact that the geometer, its operation simulated by « hand », has on several occasions produced a proof that was a complete surprise to its programmers. The nature of an intelligent program is such that unlike a conventional arithmetic computation, in which the branches are few and easily traceable, the number of conditional branches depending on the input are bewilderingly many and highly interdependent, rendering impossible any detailed attempt to trace its behavior. And of course, once learning is introduced into the program, it will constantly modify itself in a highly complex way, so that while its behavior is still in principle determined, one will become increasingly powerless to predict its response in any given case. In a very real sense, the machine's proofs will be no more or less trivial than those offered by the neophyte mathematician who is still under the influence of his professor.

One may view this machine in still another way. At any instant of time, the internal configuration of our machine is some particular state of a finite state automation. Then of the infinite number of sequences that one might ask the machine to establish as theorems, some infinite subset of these will be provable by it. At any given time, our machine represents a partial decision method over this infinite set of theorems, and this set will be richer in « interesting » theorems than a random subset of all theorems. The class of theorems considered « interesting » will determine the heuristics that control the partial decision method, and in turn, the density of interesting theorems in the set enumerated by the machine will depend on the apt choice of the heuristics. It is important to note that if even the most rudimentary learning behavior is built into the machine, its initial internal configuration will be different for each new problem presented to it, and consequently, the class of theorems decidable by the machine will be continually changing. And what is any human mathematician but a partial decision machine over some unknown class of theorems?

It is possible to approach the problem of theorem proving by machine from a rather different direction. E. W. BETH describes a method (« semantic tableaux ») for systematically constructing a counter-example for a proposed theorem if there is one, or else establishing the fact that none exists [4]. If it can be shown that a counter example cannot be constructed, an algorithm is given for converting the « closed » semantic tableaux produced into a proof of the theorem in the formal system. But the method of semantic tableaux is essentially an enumeration procedure—in this case, it is the set of individual instances of the theorem that could possibly be counter examples to the theorem that is being enumerated, and like all such procedures, the bulk of calculation required rapidly outdistances the capacity of conceivable computing machines. In order to make the procedure reasonably efficient, heuristic rules for the control of the enumeration must be introduced, and one is faced with essentially the same problem that concerns the body of this paper. The more or less anthropomorphic approach followed by the authors has the advantage that suitable heuristics are readily suggested by introspection and the methods developed are more likely to be applicable to the solution of problems in non-formal systems.

9. – The theory machine.

At various points in the preceding discussion, a line of reasoning was terminated by the comment that harder problems exist but they are outside the scope of the matter being considered. This large new class of problems and how a machine can handle them is the subject of this section. We consider now a machine that takes its environment more seriously.

The subject will be introduced by an example of a more advanced kind of geometry machine, a machine that tries to learn what kind of geometry fits the environment it finds around it. The heuristic computer is provided with an environment by the diagram computer. It looks to the environment for heuristics, for clues about what to do next. However, if it learns that some measurement contradicts something that it can prove in the syntax computer, it assumes that the measurement is in error. In other words the formal system is sacrosanct.

Now suppose that the diagram computer is replaced with another that does its drawings on the surface of a sphere. Suppose further that the priorities in the heuristic computer are readjusted so that it believes the diagram computer rather than the syntax computer when the two are in conflict. Suppose also that it is provided with the means to modify the formal system and additional heuristics to enable it to do so efficiently. It would be arranged so as to try to bring theory (the syntax computer) and experiment (the diagram computer) into harmony, and thereby discover what kind of a world it lives in. This is a theory machine.

There seems to be, in principle, no reason why a theory machine should not be fitted with the means to do experimentation, a tool room, a stockroom, and an instrument room, and told to work out the theory of something or other. In practice, there is the familiar difficulty of speed and cost. Today it is cheaper and quicker to use people to do research, but perhaps someday machines will do the research and people will merely control the doing of research. This is precisely parallel to the digging of excavations. Once people did it, but now machines do the digging and people merely control the machines. The scientist using a machine to do research would have a role analogous to that of a professor at a university directing his graduate students.

A further conjecture along this line relates to programming. A person finds it much easier to communicate a complex message to another person than to a machine. Speaking is relaxed and easy while writing a program of machine instructions is detailed and exacting. When one person listens to another he often fails to interpret some word correctly for a while but later some other words enable him to understand the earlier word. It seems as if the listener is continually generating hypotheses about what the speaker means and is continually checking these hypotheses and accepting them or rejecting them and casting about for others. In terms of human activity, theorizing is much too pretentious a word for this activity. However, from the point of view of machine design, it may be that only a theory machine will be easy for people to instruct.

The interaction between formal and heuristic procedures in a theory machine is more intricate than in a theorem machine. To determine the consequences of its present hypothesis the theory machine must use the methods of the theorem machine. Because of the different nature of the typical problems it will be solving, the theory machine must lean more heavily on semantic heuristic as a substitute for rigorous deduction. Then when it finds a discrepancy between theory and experiment it must use both rigorous deduction and heuristic procedures to modify its formal system. It is an interesting feature of such a machine that the rules for formal deduction used to modify the formal system are actually part of the formal system. This is not an unreasonable situation; it is essentially what happens when the program for a calculator causes the calculator to modify the program. However, it surely is complicated, and the complication does not end here.

The machine described so far resembles a theoretician with little or no experimental skill. Additional heuristic is required to enable the machine to select a clean experiment that will be an effective test of a theory. Contingencies will arise in the experimentation, and the machine must handle these as subproblems. In other words it must invoke this whole apparatus over again at a lower level.

The theory machine is a device that conjectures about its environment and tests its conjectures. In so doing it gains an increased understanding of what is going on. It is hoped that not only will the theory machine be able to do research, but will also be easier to communicate with than a present day automatic calculator.

10. – Summary.

In contrast to the present use of automatic calculators which outperform humans in clerical tasks, the theorem machine is advanced as a device that reasons heuristically. It is therefore able to solve harder problems, and the study of it reveals some things about the nature of problems and of machines. The essential operating principle of this kind of artificial intelligence is that it has a formal part, a syntax computer that can make deductions, and a heuristic part that can make guesses. By using the syntax computer to test the guesses made on a heuristic basis, the machine is able to get results that are beyond the scope of a purely deductive machine.

Heuristic processes can be syntactic, that is depending on the language in which the problem is stated, and on the statement in that language, or they can be semantic and depend upon an interpretation or model of the formal system, in other words, an example.

The artificial geometer is an example of a theorem machine. Geometry was chosen, not because of any inherent interest, but rather because it provides an example of a problem at the right level of difficulty that needs semantic heuristic in a major way. It is being pursued by simulation on the Type 704 Electronic Data Processing Machine.

An interesting aspect of geometry is that as taught in high school, it is not rigorous. Some facts are established by proving them and some by observing the figure (*i.e.* semantic heuristics). This is a powerful, effective method of reasoning used by people and by the artificial geometer. While it would be possible, and probably easier, to make the artificial geometer perfectly rigorous, it is more significant in the study of artificial intelligence to avoid the strictness of rigor that is a proper part of metamathematics but not efficient in mathematics.

Beyond the theorem machine is the theory machine which, by conjecturing and testing the conjectures, gains an understanding of its environment. Such a machine should be able to do research and should be easier to communicate with.

The largest obstacle to the development of useful theorem and theory machines is the problem of speed. This cannot be cured by faster components alone. The major contribution to speed must come from improved heuristic so that the machine will waste less time in fruitless endeavor. The nature of hard problems insures that the machine must waste some time on wrong hunches but the waste must be kept within bounds. The machines themselves are expected to make a major contribution to the understanding of artificial intelligence because they learn as they work, and what they learn reveals much.

* * *

In closing the authors wish to acknowledge the contributions of A. NEWELL, J. McCARTHY, M. L. MINSKY, and H. A. SIMON whose relation to the project has been indicated in the text, and to C. L. GERBERICH, J. R. HANSEN, and R. M. KRAUSE, whose technical and programming contributions are making the project possible. Professor McCARTHY in particular has been playing a continuous role as consultant to the authors.

REFERENCES

[1] A. NEWELL and H. A. SIMON: *IRE Trans.*, Vol. IT-2, 61, (September 1956).

[2] A. M. TURING: *Can a machine think?*, in *Mind* (1950).

[3] A. M. TURING: *Proc. Lond. Math. Soc.*, II, **24**, 230 (1936).

[4] E. W. BETH: *Semantic entailment and formal derivability*, in *Mededlingen der Konin klik Nederlandse Akademie van Wetenschappen, aft. Letterkunde*; Nieuive Reeks, Deel 18, no. 13. See also: A. ROBINSON: *Proving a theorem* (*as done by Man, Logician, or Machine*), in *Transcription of the Proceedings of the* 1957, *Cornell Summer Inst. of Logic* (Ithaca, N. Y.).

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 195
3° Trimestr

Questions of Linguistics.

M. HALLE
Massachusetts Institute of Technology - Cambridge, Mass.

Introduction.

We begin our consideration of the linguist's approach to the problem of speech communication by inquiring into the nature of the data that constitute the subject matter of linguistics. We want to know what kind of problems are of special interest to the linguist, for only if we understand this will we be in a position to appreciate the reasons for the ways of the linguist which frequently seem strange to the outsider.

As a first answer it might be proposed that linguistics is concerned with characterizing the class of acoustical signals which men make in speaking. The natural way of going about this would be by investigating in detail the anatomical structures in man that make it possible for him to emit this special set of signals. One would investigate the human vocal tract: the larynx, the pharynx, the nasal cavity, the mouth, the tongue, the lips, etc., and one would attempt to make statements about the motor capabilities of these organs. Once one had learned all there is to know about these physiological aspects of the problem, and, provided one knew a great deal of acoustics, one could give the desired description of the acoustical signals which such a mechanism was capable of emitting. One might further investigate the analogous mechanisms in other animals and might succeed in showing how the latter differ from those of man and how this difference accounts for the differences in the respective acoustical outputs. The results of this inquiry would explain why the acoustical signals emitted by men in speaking differ from those of other animals.

This is a very important area of study, and linguistics is vitally interested in these questions. Yet these questions do not exhaust the problems of concern to the linguist: they are but a small part of the puzzles that the linguist would like to solve. As a matter of fact, if linguistics were limited to a consideration of these problems, there would hardly be any need for a separate discipline, since all of the above problems are dealt with by physiology and acoustics.

What makes linguistics as a field of enquiry quite different from physiological acoustics is the fact that what is commonly referred to as « linguistic behavior » covers a much broader area than the acoustical properties of speech, though—as I have already said—it specifically includes the latter. Let me now describe a few of these additional problems.

We have all had the experience of hearing people speak with a foreign accent. Thus, for instance, we all know people who are physiologically normal, who yet find it difficult to distinguish sounds that we ourselves have no difficulty whatever in distinguishing. For instances no English speaker would ever confuse the words « bitch » and « beach »—not even under conditions of high noise, as G. A. Miller has shown. Yet a speaker of Russian or Italian would find it extremely difficult to keep them consistently apart. Clearly the difference in the behaviour of English and foreign speakers is not physiologically determined, because the foreigner can—when his attention is drawn to it—make the required distinction. The difference in behavior is, of course, due to the fact that English, Russian, and Italian are different languages, and that different languages use different sounds.

It may, therefore, be proposed that adult speakers have established a particular behavior pattern of their vocal organs and that this behavior pattern accounts for the observed difficulty. Differences in language may, therefore, be equated with different habitual movements of the tongue and lips and with different co-ordinations of these movements. In other words, one might conceivably explain linguistic differences on a physiological-acoustical basis, provided one allowed for some learning.

This, however, is not really an adequate explanation. Consider, for instance, the manner in which Latin is spoken by priests of different nationalities. An English-speaking priest may read mass with a sound repertory that is 100% English, and a French priest may read the same mass with a sound repertory that is 100% French. Yet there is no sense to the statement that the language of the mass is anything but Latin.

An attempt may still be made to save the view of language as a purely physiological-acoustical phenomenon by saying that, *e.g.*, the English priest uses the sounds of English with the statistics appropriate for Latin. This, however, is hardly a good solution since it raises a host of extremely difficult problems. *E.g.*, it raises the question of how it is possible to identify an utterance as English on the basis of a very short sample, which might be totally atypical. But even if this were possible, there are aspects of linguistic behavior which cannot be explained in terms of physiology and acoustics alone, regardless of the refinements introduced. I shall now give a few examples of this.

A joke quite popular among elementary school children in America is the following question and answer: *Why can't one starve in the desert? — Because of the sand which is there!* The pun is based on the fact that word boundaries are not always marked acoustically, and *sand which is* is frequently indistinguishable from *sandwiches.* Yet word boundaries are crucial in understanding the message correctly, and given enough context the speaker of English will know how to assign word boundaries even if they are not acoustically marked.

Word boundaries, moreover, are not the only boundaries which have no acoustical signal and which affect the behavior of the speaker. Consider the following ambiguities:

Old | men and women   Old men | and women

He rolled | over the carpet   He rolled over | the carpet

which are due to differences in phrase structure that are not marked acoustically.

I should like also to draw attention to another type of behavior. Every speaker of a language can perform rather elaborate transformations upon sentences. Thus, for instance, given a simple declarative sentence there is a standard way of converting it into a « yes or no » question; or given an active sentence there is a standard way for converting it into a passive. As an illustration of the latter take the sentence: *A committee opposed the change in the bill* which can be readily transformed into *The change in the bill was opposed by a committee.* In order to explain how to perform this operation we would normally use such terms as *noun phrase*, *verb phrase*, *transitive verb*, etc., in the obvious way. It is important to note, however, that here, too, there is no such thing as an acoustical signal for these categories, yet the categories are essential in order to explain the speaker's behavior.

Consider again the sentence, *The change in the bill was opposed by a committee.* The choice of *was* as against *were* is governed by the number (singular or plural) of the head of the first noun phrase; *i.e. change.* But the head of the noun phrase, which itself is a noun phrase, does not have any acoustical marker to distinguish it from other noun phrases.

It must also be noted that the head of the noun phrase governs the choice of *was* as against *were* quite independently of the number of intervening words; *e.g., The change in the bill for the promotion of the study of the mating calls of rhinoceri... etc...* ***was*** *opposed by a committee.*

Engineers and other non-linguists have usually neglected problems of the kind just surveyed, considering them either outside of their ken or relatively unimportant refinements. Linguists, on the other hand, have been keenly interested in such problems. The standard grammars of the different languages always try to do something towards solving such problems. Unfortunately the standard grammars fail to be consistent or to make clear the basis on which they operate. In what follows I shall try to present in outline a descriptive framework for language which I believe to be free of, at least, the most glaring of these failings. The exposition will begin with a review of some recent work of N. Chomsky and will go on to a discussion of the phonic aspects of language, which were not considered by Chomsky.

1. – Chomsky's analysis.

According to Chomsky every language has three distinct sets of rules which operate on three different levels. On the highest level the rules are all of the type « $X \to Y$ » where « $X \to Y$ » stands for « replace X by Y », with the restriction that not more than a single symbol can be replaced in a single rule and that $X \neq Y$.

As an illustration of these rules we can take the following (*):

Sentence → Noun Phrase + Verb Phrase + (Adverbial Phrase) (1)

Noun Phrase → (Article) + Noun + (Prepositional Phrase) (2)

Verb Phrase → Verb + (Noun Phrase) (3)

Adverbial Phrase → Adverb (4*a*)

» » → Prepositional Phrase (4*b*)

Prepositional Phrase → Preposition + Noun Phrase (5)

Article → *the* (6*a*)

» → *a* (6*b*)

Noun → *committee* (7*a*)

» → *change* (7*b*)

» → *dog* (7*c*)

» → *walk* (7*d*)

» → *result* (7*e*)

» → *bill* (7*f*)

Verb → *opposed* (8*a*)

» → *took* (8*b*)

» → *barked* (8*c*)

(*) In applying a rule the symbols in parentheses may be omitted. The rules are only partially identical with those that would appear in an actual grammar of English.

Preposition → *of* (9*a*)

» → *for* (9*b*)

» → *in* (9*c*)

The application of these rules yields a partially-ordered set of symbol sequences. We shall call each symbol sequence, a *string*, and the set of such strings generated by the rules, a *derivation*. We may illustrate the process of applying the phrase structure rules by the following derivation:

Sentence	by rule
Noun Phrase + Verb Phrase	(1)
Article + Noun + Verb Phrase	(2)
Article + Noun + Verb + Noun Phrase	(3)
Article + Noun + Verb + Article + Noun + Prepositional Phrase	(4)
Article + Noun + Verb + Article + Preposition + Noun Phrase	(5)
Article + Noun + Verb + Article + Noun + Preposition + Article + Noun	(1)
a committee opposed the change in the bill	(6*b*), (7*a*) (8*a*), (6*a*) (7*b*), (9*c*) (6*a*), (7*f*)

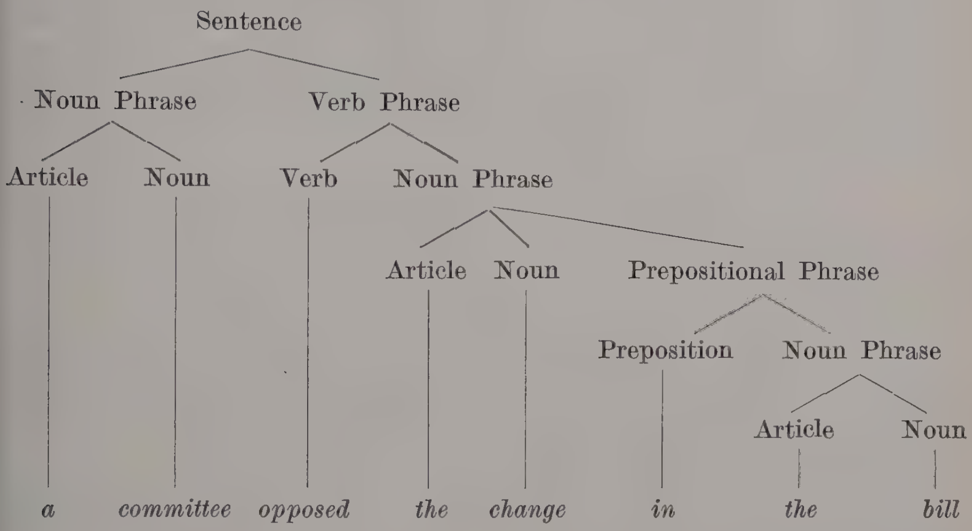
Attention must be drawn to the following facets of the grammar just presented:

1) The order of application of the rules is partly fixed owing to the fact that a given rule can be applied only if the symbol to be replaced—*i.e.*, the one appearing on the left-hand side of the rule—appears in the derivation. There must, therefore, be at least one *initial* symbol which must be supplied to the grammar from the outside and which starts things off. For the present set of rules the symbol « Sentence » will serve this function.

2) In order for the grammar to continue to operate it is necessary that instructions be provided for selecting the next rule to be applied. The instructions must be supplied from the outside. It is by exercising a choice, by selecting one rule from a set of possible alternatives that information is being transmitted. This choice must evidently be made by the user of the grammar, for only he can transmit information.

3) The grammar continues to operate as long as the string contains symbols which themselves appear on the left-hand side of one or more rules. The grammar stops operating when it has produced a string consisting of symbols which occur only on the right-hand side of the rules—e.g., *opposed* in rule (8*a*)—and hence are « irreplaceable. » We shall call these « irreplaceable » symbols, *terminal symbols*; strings consisting of terminal symbols only shall be called *terminal strings.*

It is always possible to convert a derivation into a tree like the one below.



The tree may be familiar to some readers from their school days. It represents what is commonly known as « parsing » or « diagramming » or « immediate constituent analysis » of the sentence. It contains at least a partial answer to the question of whence come the boundaries which in spite of their possible lack of acoustical correlate are nevertheless important factors in the behavior of speakers.

The restriction on the number of symbols that can be rewritten in a single rule guarantees that given a terminal string—*i.e.* a string produced by the application of the phrase-structure rules—it will be possible to discover the associated tree or trees. Since not more than one symbol can be rewritten in a single rule, every line in the derivation must have at least as many symbols as the one preceding it. Since repetitions of lines in the derivation are not admitted ($X \neq Y$), there must be a finite number of lines between the first line and the terminal string. One can, therefore, try out all one-line derivations, two-line derivations, three-line derivations, etc., until one comes upon a derivation having the desired terminal string.

Since there may be more than one derivation yielding the same terminal string, there may be more than one tree associated with a single terminal string. The fact that some terminal strings have more than one phrase structure representation accounts for the ambiguity of phrases like *old men and women*; *he rolled over the carpet*; etc.

By repeated reapplication of rules (1) and (5) endless sequences of words may be generated. This is not an oversight but rather a reflection of the fact mentioned above that language places no upper bound on the length of sentences or of constituents, although all sentences are finite in length.

We have made much of the fact that terminal strings have phrase structure. It is now necessary to point out that terminal strings are abstract representations of certain features of sentences and that actual sentences are, in fact, not terminal strings. To see this, consider the English verb. Since verbs can be in the present tense as well as in the past we introduce a rule like the following:

Verb Phrase → Verb + (Past) + (Noun Phrase) (*) (3*a*)

We would then also need rules like

oppose + Past → *opposed* (10*a*)

write + Past → *wrote* (10*b*)

have + Past → *had* (10*c*)

think + Past → *thought* (10*d*)

be + Past → *was* (10*e*)

Rule (10*a*) is within the restrictions imposed on phrase structure rules, for it requires in effect that the symbol « Past » be replaced by *-d*. The other four rules, however, violate the phrase structure constraints. *E.g.*, in (10*b*) the two symbols « *write* » and « Past » are replaced by « *wrote* » in one step, and it is impossible to achieve the same result if only a single symbol were allowed to be replaced in a single rule. Consequently, rules (10*b*) to (10*e*) are beyond the power of the phrase structure level. Since all verbs violating the phrase structure constraints belong to the so-called « strong » or « irregular » verbs of English it may be proposed that these verbs be handled as exceptions; there would then be no need to utilize more powerful devices in the grammar. We shall see, however, that the phrase structure grammar is not powerful enough to handle other, perfectly regular verbal formations in a reasonably economical fashion. The proposal to consider the « strong » verbs as exceptions is, therefore, of little practical importance.

(*) We are disregarding the problems raised by number and person.

Consider now the Verb Phrases:

had opposed	*was opposing*	*had been opposing*
had written	*was writing*	*had been writing*
had had	*was having*	*had been having*
had thought	*was thinking*	*had been thinking*
had been	*was being*	*had been being*

In order to generate the examples in the first column we should need the rule

Verb Phrase → *have* + (Past) + Verb + Perfect Participle + (Noun Phrase) (3*b*)

as well as

oppose + Perfect Participle → *opposed* (11*a*)

write + Perfect Participle → *written* (11*b*)

have + Perfect Participle → *had* (11*c*)

think + Perfect Participle → *thought* (11*d*)

be + Perfect Participle → *been* (11*e*)

In order to generate the examples of the second column we should need the following rules:

Verb Phrase → *be* + (Past) + Verb + Present Participle + (Noun Phrase) (3*c*)

and

Verb + Present Participle → Verb + *-ing* (12)

Finally in order to generate the examples in the third column we need the following additional rule:

Verb Phrase → *have* + (Past) + *be* + Perfect Participle + Verb +
+ Present Participle + (Noun Phrase) (3*d*)

This rule, however, is the sum of rules (3*a*-*c*). It is, therefore, natural to investigate whether the set of rules cannot be simplified. Examining rules (3*a*-*d*) we note the following regularities:

a) The symbol « Past » is always associated with the first element of the Verb Phrase.

b) If the Verb Phrase contains the auxiliary verb *have* the symbol « Perfect Participle » appears after the next element of the Verb Phrase.

c) If the Verb Phrase contains the auxiliary verb *be*, the symbol « Present Participle » appears after the next element of the Verb Phrase.

d) If both auxiliary verbs *have* and *be* occur in the same Verb Phrase, *have* precedes *be*.

e) The only element which must appear in the Verb Phrase is (the main) Verb.

f) The auxiliary verbs precede (the main) Verb.

The simplest way of handling these regularities is by positing the following two rules:

$$\begin{aligned}\text{Verb Phrase} \rightarrow (\text{Past}) + (\textit{have} + \text{Perfect Participle}) + \\ + (\textit{be} + \text{Present Participle}) + \text{Verb} + (\text{Noun Phrase})\end{aligned} \tag{3'}$$

and

$$\text{G} + \text{V} \rightarrow \text{V} + \text{G} \tag{Z}$$

where V stands for any specific verb (lexical morpheme) like *oppose*, *have*, *be*, *think*, etc., and G stands for a grammatical operator like « Perfect Participle, » « Past, » etc.

Rule (Z) goes clearly beyond phrase structure, for it changes the order of the symbols, and once the order of the symbols in the strings is changed, there is no longer any possibility of associating a tree with a string. We are, therefore, faced with the alternative of either maintaining the phrase structure restriction and thereby greatly complicating our description—*e.g.*, we would be forced to have four separate rules in place of the single rule (3)—or of admitting into the grammar new rules that are more powerful than those of the phrase structure level. There are various reasons why the latter alternative is to be preferred. Accordingly we establish a second grammatical level, which, following CHOMSKY, we call the *transformational level.*

It is not possible here to go into the details of the transformational level. These can be found in CHOMSKY's book *Syntatic Structures*. I should like, however, to draw attention to a few consequences of the decision to introduce the transformational rules.

Since rule (Z) must precede rules like (10) and (11), the latter together with (Z) are part of the transformational level. This makes it unnecessary to do anything special about the « strong » verbs (rules (10*b*-*d*)), since on the transformational level the prohibition against replacing more than one symbol in a single rule does not hold.

The terminal strings, the final output of the phrase-structure rules, will contain symbols of two types: lexical morphemes like *oppose*, *committee*, *of*, *the*, etc., and grammatical operators like « Past, » « Perfect Participle, » etc. This is due to the fact that at least some grammatical operators cannot be replaced by phrase structure rules; *e.g.*, « Past » is replaced in rules (10*a*-*e*), which are, however, transformational and not phrase structure rules.

The terminal string corresponding to our sample sentence is therefore represented, with some simplifications and omissions, as follows:

$$\textit{a} + \textit{committee} + \text{Past} + \textit{oppose} + \textit{the} + \textit{change} + \textit{in} + \textit{the} + \textit{bill}$$

The transformational rules operate on terminal strings *and* the trees associated with them. The notion « head of noun phrase » which we have had occasion to use in the above discussion has an obvious and simple meaning if reference is made to the tree associated with the particular noun phrase. It is a matter of considerable difficulty to give a clear meaning to this notion if one limits oneself only to the terminal string.

Up to this point we have been concerned exclusively with what might be termed abstract properties of language and we have said nothing of its acoustical features. It is now necessary to examine the relationship between the abstract entities that have been described in the preceding pages and the concrete sound waves that comprise the spoken message.

2. – Sounds of speech.

The problem with which we shall be concerned in this lecture is the manner in which the sounds of speech are to be described. In every science the choice of a descriptive framework is an extremely important matter. It is usually not enough that the description reflect the physical facts to a sufficient degree of precision. We would like to describe these facts in such a way as to open up the possibility of saying other things of interest, too. The following example illustrates this point as it may affect the linguist.

English speakers form the regular plural of nouns by adding a *sound* or *sounds* to the singular stem. They add [ɪz] if the noun ends in [s], [z], [š], [ž], [č], [ǯ], (*e.g.*, *busses*, *causes*, *bushes*, *garages*, *beaches*, *badges*); they add [s] if the noun ends in [p], [f], [t], [θ], [k], (*e.g.*, *caps*, *cuff*, *cats*, *fourths*, *backs*); and they add [z] in all other cases.

In stating this we have, however, made a number of decisions regarding the manner in which we shall describe the facts. We have spoken of individual sounds—let us henceforth call them segments—and we have attached labels to them; *e.g.*, [s], [z]. We have decided in effect to view utterances as sequences of a number of discrete entities. If we were asked why we made this decision we would surely reply that this seems to us to lead to a simple description of all kinds of facts. The questioner being a linguist in disguise might then point out that our description would be even simpler if we had a label for the segments [s], [z], [š], [ž], [č], [ǯ], and another one for the segments [p], [f], [t], [θ], [k]. But this is indeed the case if we describe the segments with the help of any of the standard phonetic frameworks: the first set consists of the noisiest sounds in the English language, variously called *hushing* and *hissing* or *strident* sounds, and the second set contains only voiceless sounds. In other words, the classification of sounds into strident and not strident (mellow), and voiced and voiceless fits well with the above facts.

We can now simplify the previous formulation in the following rules:

R. 1 If the noun ends in a strident consonant, then Plural → [ɪz].

R. 2 If a noun ends in a consonant which is voiceless, but is not strident, Plural → [s].

R. 3 In all other cases, Plural → [z].

In order to obtain simple rules we have described the utterances of English in a very special way. In particular we have regarded the utterances as consisting of sequences of discrete segments, and we have viewed the segments as simultaneous actualization of sets of attributes like voicing, stridency, consonantality, etc.

It is a well-known fact that viewed as an acoustic phenomenon speech is quasi-continuous; in many instances there is no obvious procedure for segmenting the continuous acoustic signal in a way which would correspond with the segmentation imposed by linguistic considerations. The question may, therefore, arise: in what sense can utterances be said to consist of discrete entities in sequence?

While a rigorous segmentation procedure which would show in all cases a one-to-one correspondence with the linguistic representation, may not be possible, it is possible to construct devices which produce speech by utilizing a set of discrete instructions which coincide closely with the linguistic segmentation. The devices I have in mind are of the type of the Bell Telephone Laboratories' Voder or the Haskins Laboratories' Octopus. The signal emitted by these devices is continuous speech, yet the input instructions are discrete. There is, therefore, a good sense in which utterances can be said to be made up of discrete segments.

In addition to viewing utterances as consisting of discrete segments we have also viewed the segments as simultaneous actualizations of a set of attributes. In the descriptive framework with which we will be concerned below, the number of such attributes is quite small, about 15. These 15 attributes are sufficient to characterize all segments in all languages. Since we cannot have knowledge of all languages—*e.g.*, of languages which will be spoken in the future—the preceding assertion must be understood as a statement about the nature of human language in general. It asserts in effect that human languages are phonetically much alike, that they do *not* « differ from one another without limit and in unpredictable ways. » Like all generalizations this statement can be falsified by valid counter-examples. It can, however, not be proven true with the same conclusiveness. The best that can be done is to show that the available evidence makes it very likely that the statement is true. Most important in this connection is the fact that all investigations in which large numbers of languages have been examined—from E. SIEVER's *Grundzüge der Phonetik* (1876) to TRUBETZKOY's *Grundzüge der Phonologie* (1939) and PIKE's *Phonetics* (1943)—have operated with an extremely restricted set of attributes. If this can be done with about a hundred languages from all parts of the globe, there appears good reason to believe that a not greatly enlarged catalogue of attributes will be capable of handling the remaining languages as well.

The phonetic attributes and the segments are devices in terms of which the linguist represents his data. Like descriptive parameters in other sciences, these do not always stand in a simple one-to-one relationship with the observable facts. We have already had to remark on this indirect relation in the discussion of the segmentation of the utterance. A similar situation prevails with regard to the phonetic attributes. The absence of this simple relationship, however, does not mean that there is no specific connection between the descriptive devices and the data of linguistics. In the third lecture I shall attempt to outline this relationsip.

If it is true that a small set of attributes suffices to describe the phonetic properties of all languages of the world, then it would appear quite likely that these attributes are connected with something fairly basic in man's constitution, something which is quite independent of his cultural background. Psychologists might find it rewarding to investigate the phonetic attributes; for it is not inconceivable that these attributes will prove to be very productive parameters for describing man's responses to auditory stimuli in general. It must, however, be noted that for purposes of linguistics, the lack of psychological work in this area is not fatal. For the linguist it suffices if the attributes selected yield reasonable, elegant and insightful descriptions of all relevant linguistic data.

The attributes in terms of which we shall describe the sounds of speech are due primarily to R. JAKOBSON. Following JAKOBSON, we shall call these attributes *distinctive features*. The distinctive features have been described in detail elsewhere. We shall, therefore, present here only the articulatory correlates of a few distinctive features [*].

Articulatory correlates of the distinctive features (*partial list*). [**]

1. Vocalic - nonvocalic. Single vocal cord source and absence of total occlusion in the oral cavity.
2. Consonantal - nonconsonantal. Presence of major constriction in the central path through the oral cavity.
3. Diffuse - nondiffuse. Oral cavity more constricted in front than at velum (backward flanged.)
4. Compact - noncompact. Oral cavity more constricted at velum than in front (forward flanged, horn shaped.)
5. Grave - acute. Major constriction in periphery (lips or velum) of oral cavity.
6. Nasal - nonnasal. Velum lowered.
7. Voiced - unvoiced. Vocal cords vibrating.
8. Flat - natural. Lips rounded.
9. Continuant - interrupted. No stoppage of air flow through mouth.

The first two features produce a quadri-partite division of the sounds of speech into 1) Vowels, which are vocalic and nonconsonantal; 2) Liquids, [r], [l], which are vocalic and consonantal; 3) Consonants, which are nonvocalic and consonantal; and 4) Glides, [h], [w], [j], which are nonvocalic and nonconsonantal.

Like all phonetic frameworks, the distinctive feature system is a catalogue of attributes. The distinctive feature system differs from other phonetic frameworks in that it contains only binary attributes. A segment, *e.g.*, is either voiced or voiceless, and there are no intermediate degrees of voicing ot which cognizance needs to be taken.

The question may well arise whether this is more than an empty trick, since any number of distinctions can always be expressed in terms of binary

[*] The fact that in the following list, reference is made only to the articulatory properties of speech and nothing is said about the acoustical properties, is not to be taken as an indication that the latter are somehow less important. The only reason for concentrating here exclusively on the former is that these are more readily observed without instruments. If reference were to be made to the acoustical properties of speech it would be necessary to report on experimental findings of fair complexity which would expand the present lecture beyond its allowed limits.

[**] Each feature is designated by a pair of antonymous adjectives, which, in accordance with the following convention, are used also to designate the segments: If the given description applies to a segment, it is designated by the first adjective; if the description does not apply, the segment is designated by the second adjective.

properties. All phonetic frameworks incorporate a large number of binary attributes: *e.g.*, voicing, nasality, rounding, aspiration, palatalization, etc. It is, of course, possible to replace these attributes by multi-valued properties. No one has ever shown, however, that anything is to be gained by this substitution. The replacement of multi-valued properties by binary features, on the contrary, does result in a gain.

In order to see this we shall examine the so-called *point of articulation.* The « point of articulation » is the place of maximum constriction in the oral cavity, and it has been customary to describe consonants in terms of this point. Thus, for instance, [p] is usually said to have a bilabial point of articulation, [f], a labio-dental point of articulation, [t], a dental or post-dental point of articulation, [k], a velar point of articulation, etc. No limitation is placed on the number of such points. In any given language, however, the number of separate points that need to be recognized is rather small. As a matter of fact, it can be shown that four such points suffice to describe all relevant facts in any known language. Instead of the multi-valued point of articulation dimension, the distinctive feature system contains the two features compact-noncompact and grave-acute, which distinguish the required four classes of segments: [p] is noncompact grave, [t] is noncompact acute, [c] as in *keys* is compact acute and [k] as in *cool* is compact grave.

The distinctive feature system employs less descriptive machinery than do other phonetic systems. Whereas in other systems the number of possible points of articulation is not restricted, in the distinctive feature system there are only as many different classes as are absolutely necessary. The decision to replace the point of articulation by two binary features, however, has other interesting consequences as well; *e.g.*, it makes it possible to explain in a simple manner certain linguistic changes which have puzzled linguists for a long time. One such example we shall examine in some detail.

It has been observed that when sounds change, these changes are gradual. *E.g.*, it is quite common for a voiced consonant to change into its voiceless cognate or vice versa ([v] → [f] or [k] → [g]); it is uncommon, or perhaps even unknown, for a voiceless consonant to change into a vowel ([k] ↛ [u]; [f] ↛ [a]). This observation can be conveniently expressed in terms of distinctive features as follows: a sound change rarely affects more than one feature.

In certain languages it has been found that [k] changes into [p] or vice versa. In terms of the multi-valued point of articulation this change is rather surprising, for [p] and [k] are produced with constrictions at opposite ends of the oral cavity. One might expect a change of [p] to [t] since they have adjacent points of articulation, but it seems rather curious that [p] and [k], which are articulated at such widely separated points should be confused. The distinctive feature system, however, provides a simple explanation for the puzzle. In terms of the distinctive features [p] and [k] differ in only a single feature: [p]

is noncompact and and [k] is compact. Consequently, the change of [p] into [k] is structurally quite similar to the change of a voiced consonant into its voiceless cognate.

The second difference between most standard systems and the distinctive feature system lies in the treatment of the two major classes of segments, the vowels and the consonants. In most standard systems these two classes are described in terms of features which are totally different: consonants are described in terms of the « points of articulation, » whereas vowels are described in terms of the so-called « vowel triangle. » In the distinctive feature system, on the other hand these two classes are handled by the same features: compact-noncompact, (diffuse-nondiffuse) and grave-acute. The distinctive feature system is thus more economical than other phonetic systems (*).

3. – Phonology.

Utterances are represented as sequences of distinctive feature segments. Although in many instances the latter stand in a one: one relationship with the sounds that we speak and hear, there are many instances where this relation is anything but simple. It is the major aim of the present lecture to elucidate this connection. The part of linguistics that is concerned with this problem is called *phonology*.

The phrase structure grammar, which was presented in Sect. **1**, contained rules like « Noun → *committee*, *bill*, etc. » - cf., rules (6)–(9). These rules are basically lists of all existing morphemes in the language. Our purpose in preparing a scientific description of a language is, however, not achieved if we give only an inventory of all existing morphemes; we must also describe the structural principles which underlie all existing forms. Just as syntax is not identical with an inventory of all observed sentences of a language; so phonology—*i.e.*, a description of its phonic aspects—is not identical with a list of existing morphemes.

In order to generate a specific sentence it is necessary to supply to the grammar instructions for selecting from the lists of morphemes—*i.e.*, from the morphemes appearing on the right hand side of rules (6)–(9)—the particular morphemes appearing in the sentence. Instead of using an arbitrary numerical code which tells us nothing about the phonetic structure of the morphemes, it is possible—and also more consonant with the aims of a linguistic description—to utilize for this purpose the distinctive feature representation of the mor-

(*) It is curious to note that the Hindu phoneticians had the idea of treating vowels and consonant together over 2000 years ago. Their solution differs from the one proposed here in that it classified vowels as well as consonants in terms of their points of articulation.

phemes directly. In other words, instead of instructing the grammar to select noun (7*f*), we instruct the grammar to select the noun which in its first segment has the features: nonvocalic, consonantal, noncompact, grave, voiced, etc.; in its second segment, the features: vocalic, nonconsonantal, diffuse, acute, etc.; in its third segment, the features: vocalic, consonantal, etc. Instructions of this type need not contain information about all features but only about features or feature combinations which serve to distinguish one morpheme from another. This is a very important fact since in every language only certain features or feature combinations can serve to distinguish morphemes from one another. We call these features and feature combinations *phonemic*, and we can say that in the input instructions only phonemic features or feature combinations must occur.

Languages differ also in the way they handle nonphonemic features or feature combinations. For some of the nonphonemic features there are definite rules; for others the decision is left up to the speaker who can do as he likes. *E.g.*, the feature of aspiration is nonphonemic in English; its occurrence is subject to the following conditions:

a) All segments other than the voiceless stops [k], [p], [t] are unaspirated.

b) The voiceless stops are never aspirated after [s].

c) Except after [s], voiceless stops are always aspirated before an accented vowel.

d) In all other positions, aspiration of voiceless stops is optional.

A complete grammar must obviously contain a statement of such facts, for they are of crucial importance to one who would speak the language correctly.

In addition to features like aspiration in English, which are never phonemic, there are features in every language which are phonemic, only in those segments where they occur in conjunction with certain other features, and are not phonemic in other segments. *E.g.*, in English the feature of voicing is phonemic only in the nonnasal consonants; all other segments except [h] are normally voiced, while [h] is voiceless.

So far we have dealt only with features which are nonphonemic regardless of neighboring segments. There are also cases where features are nonphonemic because they occur in the vicinity of certain other segments.

As an example we might take the segment sequences at the beginning of English words. It will be recalled that the features vocalic-nonvocalic and consonantal-nonconsonantal distinguish four classes of segments: Vowels, symbolized here by V, are vocalic and nonconsonantal; Consonants, symbolized by C, are nonvocalic and consonantal; Liquids [r], [l], symbolized by L, are vocalic and consonantal; the Glide [h], symbolized by H, is nonvocalic and consonantal (*). We shall be concerned solely with restrictions on these four classes; all further restrictions within the classes are disregarded here.

English morphemes can begin only with V, CV, LV, HV, CCV, CLV, and CCLV: *e.g.*, *odd*, *do*, *rue*, *who*, *stew*, *clew*, *screw*. A number of sequences are not admitted initially; *e.g.*, LCV, HLV. These constraints are reflected in the following three rules which are part of the grammar of English:

Rule MS1: If a morpheme begins with a consonant followed by a nonvocalic segment, the latter is also consonantal.

Rule MS2: If a morpheme begins with a sequence of two consonants, the third segment in the sequence is vocalic.

Rule MS3: If between the beginning of a morpheme and a liquid or a glide no vowel intervenes, the segment following the liquid or the glide is a vowel.

These rules enable us to specify uniquely a number of features in certain segment sequences; *e.g.*,

vocalic	—	—		
consonantal	+		+	

is converted by rules MS1, 2 and 3 into

vocalic	—	—	+	+
consonantal	+	+	+	—

which stands for a sequence CCLV: *e.g.*, *straw*.

The MS rules are partially ordered. If the order is not imposed they will have to be given in a much more complex form. Let us now introduce the convention that whenever a feature is not specified in a segment, a zero shall be written in the appropriate column and row. We shall say, therefore, that a zero stands for an unspecified feature, and a plus or a minus, for a specified feature. In terms of this convention the sequence of columns representing the different morphemes—*i.e.*, the input instructions for phrase structure rules (6)–(9)—will contain many zeros; indeed as many zeros as are compatible with attaining the aims of the grammar.

We define an order-relation between segment-types: We shall say that

(*) We consider the semivowels [j] as in *you* and [w] as in *woo* to be positional variants of the vowels [i] and [u], respectively.

segment-type A is « contained » in segment-type B, if and only if the following two conditions are satisfied: 1) all specified features of A are found with the identical values (the same pluses and minuses) in B; and 2) at least one feature specified in B is unspecified (has a zero) in A. The set of all elements not « contained » in any other element is called the *set of maximal segmenttypes.*

Examples:

		A	B	C
Feature	F1	+	—	+
	F2	0	+	—

A is « contained » in C. The set of maximal segment-types is {B, C}.

		A	B	C
Feature	F1	+	—	0
	F2	0	+	—

all segment-types are maximal.

It has often been observed in linguistics that the primary function of the phonemes of a language is to distinguish one morpheme from another. It is, therefore, natural to require that the set of phonemes of a language be a set of maximal segment-types. In other words, given any two phonemes of a language, it must be the case that for at least one feature, one phoneme has a plus where the other phoneme has a minus, or vice versa.

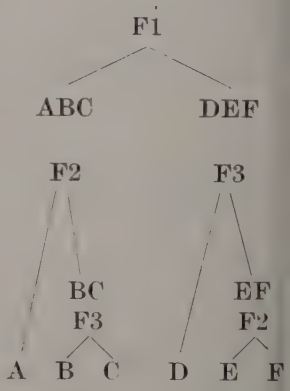
Each specified feature in a segment represents a piece of information that must be provided in the input instructions. If our grammar is a realistic picture of the language, then this information must be supplied by the speaker. Since we speak quite rapidly—at a rate which may be as high as 20 segments per second—it is only reasonable to assume that the number of specified features in the input instructions is consistently kept at a minimum. One way of approaching this desideratum is by minimizing the number of specified features per phoneme. It can be shown that if this condition is imposed on a set of maximal segment-types, it will be possible to map into a branching diagram the matrix representing the set of segment-types, in such a way that if to each node a particular feature is assigned, then each path through the diagram beginning at the initial node and ending at the end points of the branching diagram represents a phoneme.

In order to see what is involved consider the following sets of maximal segment-types.

		A	B	C
Feature	F1	+	—	0
	F2	0	+	—
	F3	—	0	+

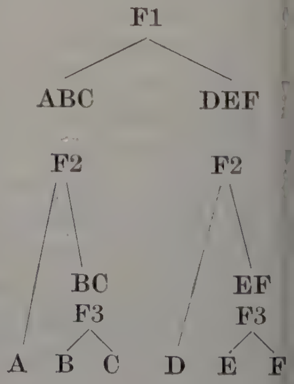
This set of maximal segment-types is not mappable into a branching diagram.

		A	B	C	D	E	F
Feature	F1	—	—	—	+	+	+
	F2	—	+	+	0	—	+
	F3	0	—	+	—	+	+



Note that in the left branch of this branching diagram, F2 precedes F3, while in the right branch the inverse order obtains. Without this reversal in the order of the features, the above set of maximal segment-types is not mappable into a branching diagram.

		A	B	C	D	E	F
Feature	F1	—	—	—	+	+	+
	F2	—	+	+	—	+	+
	F3	0	—	+	0	—	+



This set of maximal segment-types can be mapped into a branching diagram with a unique ordering of the features.

The possibility of mapping a distinctive feature matrix into a branching diagram hinges upon the existence in the matrix of at least one feature for which there are no zeros. This feature, which must be assigned to the first node, subdivides the segment-types into two classes. The next two nodes must be assigned to features which have no zeros for any of the segments in the two sub-classes. These may be the same or different features. The same procedure must again be possible with regard to the segments in each of the four sub-classes established by the former features; etc. When a sub-class contains a single segment-type, the segment-type is fully specified, and the path through the branching diagram represents exactly its distinctive feature composition. The two conditions establish a hierarchy among the features. This hierarchy, however, need not be complete. For instance, when there are in the matrix two features which contain no zeros, there is no reason to put one feature before the other; any order will be satisfactory. Partial ordering of features for different reasons is illustrated in the second example above.

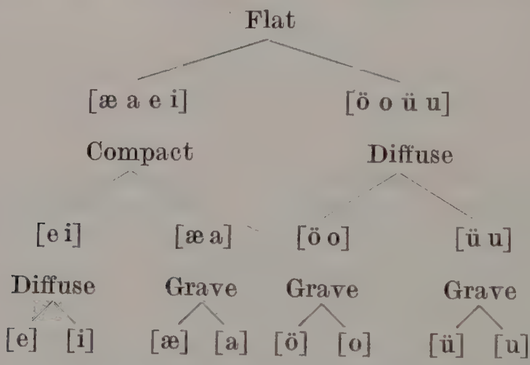
The hierarchy of features established by the two formal conditions imposed on phonemes provides an explanation for a number of observations made by linguists. It accounts, *e.g.*, for the intuition that the distinction between vowels and consonants is somehow more crucial to the phonological system than the distinction between accented and unaccented vowels, or between stops and continuants. Since in all phonological systems it happens to be the case that the features vocalic-nonvocalic and consonantal-nonconsonantal must precede all other features, it is quite natural that the segment classes established by these two features should be felt to be more central than other classifications of segments.

An interesting result of a different sort is obtained in the case of the Finnish vowel system. Finnish has the eight vowel phonemes which can be characterized by means of the following distinctive feature matrix.

	[æ]	[a]	[e]	[ö]	[o]	[i]	[ü]	[u]
flat	—	—	—	+	+	—	+	+
compact	+	+	—	—	—	—	—	—
diffuse	—	—	—	—	—	+	+	+
grave	—	+	—	—	+	—	—	+

Since, however, it is necessary to minimize the number of specified features per segment, we replace certain specified features by zeros as follows:

	[æ]	[a]	[e]	[ö]	[o]	[i]	[ü]	[u]
flat	−	−	−	+	+	−	+	+
compact	+	+	−	0	0	−	0	0
diffuse	0	0	−	−	−	+	+	+
grave-acute	−	+	0	−	+	0	−	+



This replacement of specified features by zero has, however, an interesting parallel. Finnish is one of the languages which possess *vowel harmony*; *i.e.*, there is a restriction on what vowels can occur in a single word. In the case of Finnish, a word can contain as election either from the set [æ, ö, ü, e, i] or from the set [a, o, u, e, i]. The minimal distinctive feature matrix provides us with a very elegant formula for the description of these facts; *i.e.*, a Finnish word cannot contain both grave and acute vowels. The formula holds only for the abstract representation of the phonemes as it is embodied in the matrix, for physically speaking [e] and [i] are both acute. In the construction of the Finnish word, these two phonemes, however, do not behave like other acute vowels. The formal requirements imposed on phonemes force us to treat [e] and [i] as vowels which are neutral with regard to the feature grave-acute, and indeed this is how these phonemes appear to be treated by the language.

The reasons advanced for reducing the number of specified features in the input instructions do not hold only in the case of phonemes. As we have seen in the discussion of the segment-sequences that are admitted at the beginning of an English word, under certain conditions not all features which must normally be specified in a phoneme serve to distinguish one morpheme from another. We have, however, not required that the input instructions consist entirely of phonemes. We can now take advantage of this and leave unspecified in the input instructions all features that are not phonemic. The rules of the grammar will insure in such cases that unspecified features are specified so as to yield the correct phonetic consequences; *i.e.*, possible English utterances.

The question which we have as yet not discussed is at what point in the grammar must we place the various rules that reflect the constraints on feature combinations. At first sight it may appear desirable to place all of them at the end, after the operation of the transformational rules, since it is only at the end of the transformations that all grammatical operators—*i.e.*, symbols like « Past, » « Plural, » etc.—are converted into features or feature segments. If we were to apply the above rules before the transformations it would be necessary either to apply the same rules again, in order to handle those feature segments that were introduced by the transformational rules, or to specify many more features in the output of the transformational rules. I shall now attempt to present reasons why it is necessary to apply some rules reflecting constraints on feature combinations before the transformations.

Since it is always possible to add new words to the language the lists of morphemes must not be considered closed. The rules which reflect the constraints on feature combinations do not enable us to develop a procedure for discovering the most economical distinctive feature representation for every morpheme; this can be found only by repeated trial and error. Consequently, it is not possible to predict a priori what types of distinctive feature columns will appear in the representations of the different morphemes, for it is conceivable that a new morpheme to be introduced in the future will require for its most economical representation a distinctive feature column that is not otherwise found in the language.

The above fact has important consequences for the construction of the grammar. We have just said in effect that we do not have a way for determining what distinctive feature columns (segment-types) will appear in the terminal strings after the application of the phrase structure rules. In many languages—though perhaps not in all languages—there are certain transformational rules which require that certain features be specified. As an example consider the plural of the English noun « straw » [str'ɔ]. As was shown at the beginning of this lecture the features vocalic-nonvocalic and consonantal-nonconsonantal would be represented in this morpheme as follows:

vocalic	—	—	0	0
consonantal	+	0	+	0

In other words, in the input instruction there would be no statement regarding the nature of the last segment. In order to select the correct plural ending for this noun, however, it is necessary to know its last segment (*). This information is contained in the rules reflecting the constraints on feature combinations; *i.e.* in rules MS1, MS2, MS3. It is necessary, therefore, to apply these rules before the rule forming the plural of nouns, or more generally before all transformational rules. I believe that the dividing line between the rules that have to be applied before the transformations—let us call them the *morpheme structure rules*—and those that have to be applied after the transformations—let us call these the *phonological rules*—can be drawn by requiring that the application of the morpheme structure rules result in segment-types which are specified to a point where the entire set of segment-types is mappable into a branching diagram in which each segment-type is represented by a distinct path through the diagram, all paths beginning at the initial node, but not necessarily ending in an end point. In other words, at this point the segment-types admitted in the representation are either phonemes or segment-types which are « contained in » phonemes. We shall call the latter segment types *archiphonemes*. Since, however, the entire set must be mappable into a branching diagram a feature specified in a phoneme can remain unspecified in an archiphoneme only if all features below it in the hierarchy established by the branching diagram also remain unspecified.

Since the morpheme structure rules must be applied before the transformations, it is natural to include them in the phrase structure level rather than set up a separate linguistic level containing just these rules. The MS rules must, therefore, be of the same structure as other phrase structure rules; they must, *e.g.*, not violate the restriction against rewriting more than one symbol in a single rule. They can not result, therefore, in the elimination of entire segments from the representation. Such rules, which are necessary in certain instances, will have to be included in another part of the grammar.

All remaining rules dealing with constraints on feature combinations are to be applied after the transformations. Since these rules differ from the transformations in two significant respects—namely, all the rules are obligatory; *i.e.*, require no external instructions to be put into operation; and the rules do not require reference to other, earlier strings in the derivation—it is simplest to set up a special linguistic level containing only these rules. We call this third linguistic level the *phonological level*. The rules of the phonological level complete the specification of the phonetic properties of the utterance in so far as these are governed by the rules of the language. Phonetic properties whose actualization is left to the free will of the speaker are not specified by these or any other rules. They are beyond the purview of the science of linguistics.

(*) The rule governing the selection of the plural endings in English is stated at the beginning of Sect. 2.

* * *

I want to express my gratitude to my colleagues Y. Bar-Hillel and M.P. Schützenberger for many hours of discussion and instruction, hours which they might well have spent more enjoyably.

This work was supported in part by the U.S. Army (Signal Corps), the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U.S. Navy (Office of Naval Research), and the National Science Foundation.

BIBLIOGRAPHY

N. Chomsky: *Syntactic Structures* (The Hague, 1957).

R. Jakobson, C. G. M. Fant and M. Halle: *Preliminaries to Speech Analysis* (M.I.T. Acoustics Lab. Report 13, 1952).

R. Jakobson and M. Halle: *Fundamentals of Language* (The Hague, 1956).

M. Halle: *The Sound Pattern of Russian* (The Hague, in press).

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3° Trimestre

Statistical Macro-Linguistics.

B. Mandelbrot

Institut de Mathématiques - Université de Lille

It is well-known, how a simple and economic theory may transform an empirical law from something quite amazing and difficult to believe, into something almost obvious and even trivial. It seems that such will be the final fate of certain laws of linguistics; the relationship between rank and frequency for natural words, and the relationship between species and genera in natural taxonomies. Let us recall that these laws were discovered by J. B. Estoup and J. C. Willis, respectively, but were made well known by the publications of G. K. Zipf [1]. The author's models for these results were published since 1951, and their final form was given in 1957. Although these theories are essentially very simple, we have not yet found a way of developing them fully in a few pages. We shall therefore limit ourselves in this Note to a bare outline of the theory of the frequency distribution for natural words.

The first main tool of the theory is the following relationship:

$$C = -\log_2 p \, ,$$

where p is the probability of occurrence of some signal in a message and C is the «cost» of transmitting this signal in some optimal binary code. This relationship is extremely familiar in information theory and may be obtained under a wide variety of definitions of optimality; we shall not attempt here to reduce this relationship to more fundamental concepts. Further, we shall not restrict ourselves to binary codes, and shall write:

$$\beta C = -\log_e p \, , \tag{1}$$

where β is a factor which depends upon the scale chosen for C.

Let us apply the relationship (1) to the words of natural language. Each word will be labelled by the rank, which it occupies in a list of all words, arranged by order of decreasing probability in a given text: that is, $r=1$ designates the most frequent word, $r=2$, the second most frequent, etc.; the number of words more frequent than a word of frequency p will be $r(p)-1$. Then, the empirical result is that for words other than the most frequent ones (large r) one has, whichever the language in which a test was written:

$$p(r) = Pr^{-B}, \tag{2}$$

where P and B are some constants. The relationship (1) then becomes:

$$\beta C = -\log P + B \log r,$$

$$\log r = \frac{\log P}{B} + \frac{\beta}{B} C = \log K + \beta' C,$$

(by definition of K and β'); finally,

$$r = K \exp[\beta' C], \tag{3}$$

An « explanation » of the law of Zipf requires an interpretation of the « cost » of coding a word and a model for the structure of the word, which together would lead to (3). One reasonable interpretation of « cost » would be the number of letters required for the code. It turns out actually that this interpretation cannot be carried to the end, and one must rather think of the cost as being something like the time required to read a word [2]. However we shall sketch a theory based upon the identification of cost to (essentially) the number of letters. The second step is the choice of the rule of formation of words: in the present model, one will assume that a word is any sequence of letters contained between successive occurrences of some additional improper letter, the « space ».

It is then reasonable to interpret cost as being equal to the number of proper letters, plus the cost C_0 of the improper letter « space ». Let there be M different proper letters. Then

there is	1	word	of cost C_0
there are	M	words	of cost C_0+1
there are	M^2	words	of cost C_0+2, etc.

Adding, one finds that

$$\text{there are } \frac{M^n-1}{M-1} \text{ words of cost less than } C = C_0 + n.$$

For large n, this gives

$$r = K' M^{\sigma - \sigma_0} \underset{\cdot}{=} K \exp\left[C \log M \right]$$

which is of the form required to explain Zipf's data on word frequency for large r.

It is unfortunate that the simplest case above cannot be carried out to further steps without some difficulties. However, it turns out that the same result (3) can be obtained under wider, more realistic and mathematically more convenient conditions, as long as *a word is a sequence of letters contained between two successive spaces, as long as there is « little interaction » between successive proper letters, and as soon as one can justify* (1).

A closer examination of the cost of coding for small values of r suggests the following improvement of the law (2), valid for all r,

$$p(r) = (B - 1) V^{B-1} (r + V)^{-B} ,$$

where V is a second coefficient. This further approximation turns out to be experimentally excellent (*).

(*) Other explanations may eventually be given of the rank frequency relation (2). However, the explanation suggested by H. A. Simon is certainly incorrect. See our *Note on a class of skew distribution functions*, in *Information and Control*, **2**, 90 (1959).

REFERENCES

[1] G. K. Zipf: *Human Behavior and the Principle of Least Effort* (Cambridge, Mass. 1949).

[2] B. Mandelbrot: *Linguistique statistique macroscopique*, I. One of the three essays in the book *Logique, langage et théorie de l'information*, by L. Apostel, B. Mandelbrot and A. Morf (Paris, 1957), and B. Mandelbrot: *Linguistique statistique macroscopique*, II. A report of the Institut de Statistique de l'Université de Paris (Paris, 1957).

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3° Trimestre

Morphology of Nerve Nets (*).

V. Braitenberg

Scuola di Perfezionamento in Fisica Teorica e Nucleare, Sezione di Cibernetica - Napoli.

1. – Introduction.

Neuroanatomy is what is left when all the problems about the neuron are solved. Most of these problems, however, are still unsolved and much confusion has been generated through the assumption of clear-cut findings where only hypotheses where intended. The diagram below (the McCulloch and Pitts abstraction) stands for the logical expression (in Hilbert and Ackermann notation):

$$\bar{E}\, v\, (AB\bar{D})\, v\, (AC\bar{D})\, v\, (BC\bar{D})\,.$$

It has but a very tenuous relation to neurons. Nobody has ever seen, presumably, the complete anatomical structure representing a junction such as (*A*-*E* or *D*-*E* in the case of two real nerve cells and nobody knows the exact difference between junctions of the excitatory (*A*-*E*, *B*-*E*, *C*-*E*) and of the inhibitory kind (*D*-*E*). Also, it is very doubtful whether chains of events in a nervous system can always be expressed by enumerating neurons which in an « all or non » fashion fired in succession. Ional changes taking place in limited portions of the neurons, and such changes as cannot be described in either of the categories of the « firing » or « not firing » of a neuron may play a much more remarkable role than would appear from the first approximation models. Thus it is not at all sure that the two branches *E′* and *E″* in Fig. 1

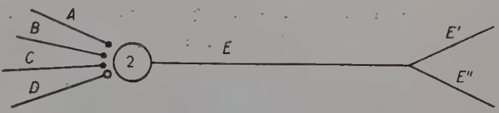


Fig. 1.

(*) The present research has been sponsored in part by the European Office of the Air Research and Development Command, U.S. Air Force.

will always « fire » when the axon E fires etc. There are no experiments to indicate the exact extent to which neurons influence each other outside of the « official » junctions. Whether activity in dendrites can produce activity in neighbouring dendrites directly and what sort of changes mediate this interaction, is just as obscure as whether similar interactions occur among axons in a parallel bundle.

2. – Methods.

Evidence about the texture of the brain is derived from the following fundamental methods of investigation:

a) *Fiber tracing*: The fact that the nervous system is largely built out of long fibres was obvious to old anatomists who investigated slightly macerated brains with the aid of wooden sticks and skilful dissection. Various phenomena which proved to remain limited to fibers or fiber bundles without spreading at right angles to the direction of the fibers helped to individuate *pathways*. Such phenomena were the degeneration *i.e.* anatomical disrupture of entire fibers following the experimental destruction of parts of the fibers, and the possibility of recording localized electric potentials after localized stimulation as evidence of existing fiber connections between the site of recording and the site of stimulation.

Theories elaborated by workers who used these techniques are sustained by the hopeful assumption that the knowledge of the connections between input stations, elaborating stations and output stations would in itself provide an answer to most problems. The aspects completely neglected are those regarding the junctions, the conditions for the propagation of signals in knots, *i.e.* places where fibers bifurcate or are confluent.

b) *Golgi method*: A very odd chemical reaction between a chromate solution, with which pieces of brain have been impregnated, and silver nitrate, into which these pieces are subsequently placed, may produce nuclei of precipitate inside a cell which grow continuously into all its ramifications without ever trespassing onto the outside or onto neighbouring cells. This produces an unpredictable selection of very few neurons out of a large neuron population, with the unestimable advantage of a « histochemical dissection » of the complete ramifications of the cell, which would of course remain inextricable if all neurons of the net were stained.

The theoretical bias produced by this view of the nerve nets was an overestimation of the importance of minute differences of nerve cell shape, while the statistical distribution of the different types was largely neglected. The neuron theory, with the dogmatic assumption of the *unit* character of nerve cells was another outcome.

c) *Aniline dyes*: It is possible to stain the « cell bodies » of neurons selectively without staining any of the long branches and fibers, thus likening neurons to other cells of the organism. This technique has the advantage of completeness (it does not leave out any cells) and is the basis of cell counts, since it is obviously easier to count small distinct dots than closely packed stars or brushes. The importance of this method has been grossly overestimated, and the quotation « nil bonum e Nissl » becomes understandable in view of the fact that these methods (the « Nissl methods ») leave us completely in the dark both about the junctions and the pathways themselves.

d) *Fiber staining*: Certain staining methods reveal remarkable local differences in different feltworks of nerve fibers. The intensity of the stain proves to coincide generally with the presence of long and thick fibers, whilst an equally dense feltwork of thin and short fibers stains very lightly. This phenomenon is due to a fatty substance, myelin, which covers the thicker fibers and performs with all probability the function of an insulator. The study of nerve nets with the aid of this method has not given all the results it may yet give. Under the general assumption that contacts between fibers are made everywhere except in the insulated portions, the presence and the orientation of long insulated fast conducting (see below, Sect. **3**) segments reveals the most important exceptions to the randomness of a net.

3. – The basic element.

In view of the uncertainty of the translation of the neurological reality into logical diagrams (Sect. **1**) the basic element should be carefully defined and should not be burdened with too many unproven assumptions.

a) The basic element of nerve nets is the *insulated nerve fiber*. It varies in length between 10 μm and over 1 m, and in thickness between about 1/10 μm and over 20 μm.

The operation performed by such a fiber is the transmission of an event from one end to the other without interference with other events in other fibers.

b) The *direction* of transmission is fixed for each fiber.

c) The *velocity* of the transmission varies between a few cm/s and over 100 m/s. It is fixed for each fiber and varies directly with the thickness of the fiber (the thicker the faster).

d) The transmission of an event is a consequence of the occurrence of « excitation » within a region, called *pickup-field*, associated with each fiber.

e) The transmission of an event produces « excitation » within a region, called *field of excitation*, associated with each fiber. Field of excitation and pickup field of the same or of different fibers may be overlapping.

The phenomena called « event » and « excitation » are complex physicochemical changes. It is convenient to treat the event which is transmitted in nerve fibers as a binary signal and to relegate to the intervenient excitation those effects which appear as monotonic (« facilitation ») or non monotonic (« inhibition ») functions of the number of active fibers.

The places where the quantity excitation determines the transmission or not transmission of an event in fibers are the *knots* of a net. Note that according to the definitions given the knots include, besides the « synaptic junctions » between neurons, also branching points of axons (which are generally uninsulated) and, generally, uninsulated segments. In praxis (Sect. **5** *b*, *c*, *d*) we shall have to revert to the conventional identification of fibers with « neurons » due to the lack of precise physiological data on the interactions outside of the synaptic regions.

4. – Some simple organs composed of fibers.

a) *Bundles.* Large parts of all brains are composed of bundles of parallel fibers. The number of fibers in a bundle is mostly of the order 10^3 up to 10^7, smaller bundles ($10 \div 100$ fibers) being rare and larger assemblies being mostly composites in no way describable as uniform organs (example, white matter of the hemispheres).

All fibers in a bundle transmit events in the same direction. If fibers of opposite direction are found in what appears macroscopically as one bundle, we prefer to speak of two bundles.

The operation performed by a bundle of parallel fibers is the translation of patterns of excitation and the production of a delay.

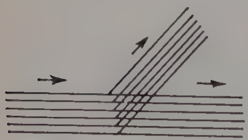


Fig. 2.

b) *Branching bundles.* Microscopical examination of fiber masses in the brain reveals places where all fibers of a bundle bifurcate, giving rise to two secondary bundles. When the direction of the fibers is known, it is always apparent that only one of the three branches of such a knot is *afferent*, *i.e.* conducts towards the knot (Fig. 2).

In the simplest case an organ composed of parallel, branching fibers performs the operations of translation, multiplication and delay of patterns of excitation.

In some cases such bifurcations will however be equivalent to the endpoints of fibers in the sense defined in Sect. **3**. In other words, excitation produced by neighbouring fibers may determine the transmission or not transmission of an event beyond a bifurcation. The classical example of this is given by the experiments of BARRON and MATHEWS (1935).

c) *Loops or delay organs.* If we eliminate the factor « translation » from the operation of a bundle by bending it into a closed ring and bringing the two endpoints of each fiber together in close proximity (Fig. 3) we obtain an organ capable of producing a fixed delay in a pattern of excitation. There is no evidence that such an organ exists in natural brains.

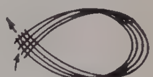


Fig. 3.

d) *Tapering bundles.* The simplest relation between a bundle of fibers and some other organ (see Sect. **5**, the griseum) is that of a one to one correspondence of fibers and subdivisions of that organ. The fibers may be afferent or efferent or mixed. The anatomical expression of this relation is a tapering bundle (Fig. 4), in the simplest case of a one dimensional array, a linear fall etc. Numerous examples could be quoted (« fimbria of the hippocampus », « optic fibers in lower vertebrate tectum », « bundles in the lateral thalamus » etc.).

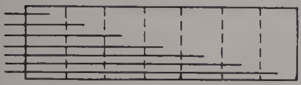


Fig. 4.

e) *Bilaterally tapering bundles* on both sides of a plane of mirror symmetry in the nervous system (the median sagittal plane in most animals) indicate the presence of a so-called *commissure*, *i.e.* of a one to one connection via fibers of symmetrical points of the two halves of the nervous system (Fig. 5). This is a frequent pattern in vertebrate brains.

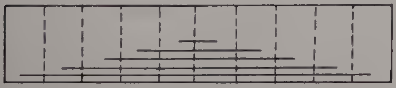


Fig. 5.

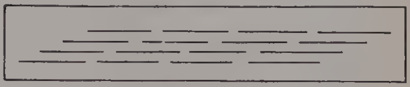


Fig. 6.

f) Not strictly speaking a bundle, but a system of fibers connecting (in the one dimensional array) all subdivisions of an organ a certain distance apart (Fig. 6) would have uniform thickness in its middle part (provided that the length of the total array is more than twice that distance) and would be seen tapering at both ends. There are examples for this in the cerebral cortex (horizontal fibers in layer IV of the striate area).

g) A system of fibers connecting each subdivision with every other subdivision of an organ, in the one dimensional array would have the shape given by $ni - n^2$ (number of fibers in the n-th of i subdivisions) (Fig. 7). Examples of this may also possibly be found in certain arrays of fibers

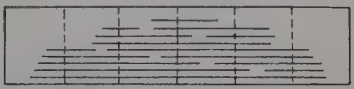


Fig. 7.

within the cerebral cortex (example: inner stria of Baillarger in the primary acoustic area on Heschl's convolution).

h) If two halves of a surface are connected with the corresponding halves of another, parallel surface through two *crossed bundles*, each composed of parallel fibers, the awkward splitting of the pattern of excitation which is produced by this arrangement becomes understandable only in connection with the analogous splitting produced by the two lenses of two separate eyes whose visual fields are not overlapping. The crossing of the optic nerves annuls the disrupture of continuity produced by the two separate optic receivers. The crossing of the vast majority of long, bilaterally symmetrical bundles in vertebrate brains is said to be secondary to the crossing of the optic connections.

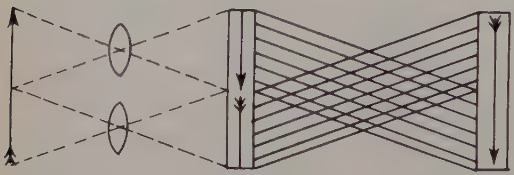


Fig. 8.

i) Bundles composed of fibers of *different thickness* and therefore different velocity of conduction (Sect. **3**-*c*) produce a temporal scattering of patterns of excitation transmitted. This could be an essential prerequisite for many mechanisms which can be theoretically postulated. Again, as in the case of loops of fibers as delay organs (Sect. **4**-*c*) we are ignorant about the physiological significance of such an arrangement. The many existing collections of fibers of varying thickness may also be interpreted as mixtures of different bundles each with uniform fiber thickness.

5. – The griseum.

A survey of existing nerve structures shows the general rule that, as insulated fibers tend to conglomerate in parallel bundles, the non insulated extremes of fibers tend to lie in delimited regions. This allows the extremes of many fibers to come into close non-insulated propinquity. A general distinction of two types of regions or of « substance » within the brain can be made, the « white substance » or *album* containing insulated fibers, and the « grey substance » or *griseum* containing non insulated extremes of fibers. The difference in shade expressed in the two denominations stems from the greater amount of whitish fatty insulating material (« myelin ») in the album.

The griseum is very varied in its fine structure and may be classified in various ways. We abstract some general characteristics concerning the relation between fibers and griseum on one hand, and some characteristics of symmetry on the other in order to arrive at a rough classification of the griseum.

a) *The ratio grey/white.* Under the assumption that the structure of a grey organ corresponds in coarseness to the coarseness of the fiber bundles connected with it, *i.e.* that the density of points capable of distinct states in a cross-section of fiber bundle is of the order of the density of such functionally distinct points in the griseum, the volume of a grey organ compared to the cross-sectional surface of all fiber bundles connected with it represents a measure of the complexity of the operations performed. This measure, which is available for many grey organs existing in nature, is generally referred to as the « grey/white index ».

b) *Total convergence or divergence.* Considering a grey organ with distinct incoming and outgoing fibers, the relative size of the two bundles may indicate convergence or divergence. A complete classification machine (UTTLEY), classifying all possible constellations of activity in n input fibers would show great divergence ($n:2^n$). A scanning device capable of transforming some spatial array of excitation into a temporal sequence would show convergence etc.

c) The influence which incoming fibers exert on outgoing fibers of the same griseum may be direct, or mediated through other fibers (internal fibers). These interactions vary according to the size, shape and relative position of pickup and excitation fields of the various fibers involved.

Pickup fields. Although for any specific fiber the locus of the points in which excitation may determine an event can be described anatomically as the surface of the corresponding « somatodendritic tree », the distribution of pickup points provided by such trees is more conveniently described as a density field around the endpoint of the fiber. This field may be approximated by some exponential function of distance as in the simplified example considered by SHOLL (1953). The size and outline of such pickup fields can be obtained through direct histological observation of the griseum. The size varies for different grey organs and for different types of fibers, the variation (in human brains) being contained between a diameter of about 10 μm and that of about 3 mm.

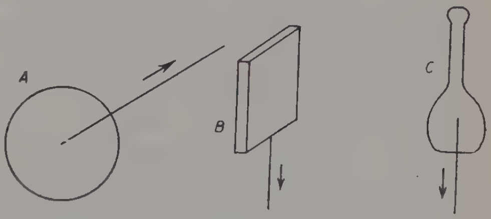


Fig. 9.

The larger pickup fields may be overlapping with as many as $10^3 \div 10^4$ other pickup fields. Some (Purkinje cells of the cerebellar cortex) never overlap.

The outline (Fig. 9) is spherical (A), or that of a flat box (B) (Purkinje cells), or, most frequently in the cerebral cortex, comparable to a very elongated pear, or a round bottle with a long neck (C).

d) *Fields of excitation.* If we take the terminal arborization of an axon as the anatomical expression of the corresponding field of excitation, we find that the outline and size of these fields is even more varied in different grey organs than the outline and size of pickup fields. The complete field of excitation corresponding to one fiber may be composed of several separate regions, due to the possibility of branching of fibers (Sect. **4**-*b*). Some extreme variations are the following (Fig. 10):

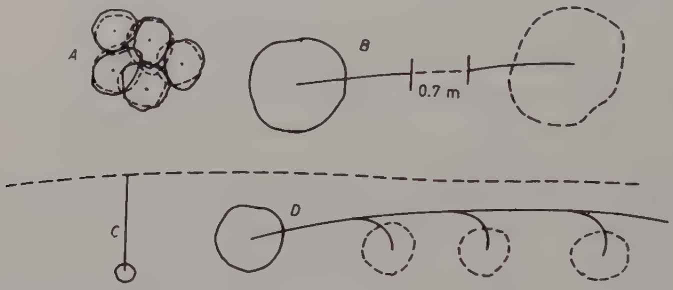


Fig. 10.

Field of excitation coinciding nearly with pickup field of the same fiber (example: granula of the cerebral cortex) (*A*);

field of excitation far removed from the corresponding pickup field (example: Betz cells of the cerebral cortex) (*B*);

very dense branching within a narrow region (example: « specific afferents », *i.e.* sensory incoming fibers of the cerebral cortex);

diffuse branching over large regions (example: reticulo-cortical neurons);

long straight unbranched and uninsulated fibers, producing considerable temporal scattering in a linear field of excitation (example: « parallel fibers » of the cerebellar cortex) (*C*);

chains of separate fields of excitation (example: « basket cells » of the cerebellar cortex) (*D*).

e) *Internal fibers.* An internal fiber of a grey organ is one that remains entirely within that organ, or in other words, a fiber which is neither afferent nor efferent. The number of such internal fibers varies greatly in different regions of the griseum. They may be missing altogether (examples: « dentate nucleus of the cerebellum », many grey organs of the « brain stem ») or correspond roughly in number to the number of afferent or of efferent fibers (example: « retina ») or outnumber greatly the external fibers (example: cerebellar cortex, where the number of « granula » is about 3000 times the number of « Purkinje cells », the only efferent fibers). The preponderance of the internal fibers is evidently related to the grey white index (Sect. **5**-*a*).

The system of internal fibers may be arranged anatomically « in series » with the external fibers (*e.g.* retina). In the simplest case this arrangement would simply iterate the transformation which a pattern of excitation undergoes in passing through the region of interlocking pickup fields and fields of excitation.

In some instances (example: « granular layer » of the cerebral cortex) the internal fibers form a closed assembly of very small and very numerous elements of the type of Fig. 10*A*, among the much larger external fibers of the same grey organ. It is possible that this provides for a « quasi analogue » mechanism capable of smoothing out the more « digital » operation of the larger elements.

Finally there exists, in many grey organs and particularly in the cerebral cortex, a dense population of relatively thick (fast conducting) long internal fibers, spanning thousands of the above mentioned granula. This population varies in density and arrangement in different regions of the same grey organ (sometimes showing the patterns described in Sect. **4**-*e*,*f*,*g*) and promises to give important clues about the variations of the structure of the net and their relation to the operations performed.

6 – Symmetry of the griseum.

The fine structure of the nerve net may be uniform throughout a grey organ, or may vary from point to point. Moreover a statistics of the meshes of the net may or may not reveal directional differences. Accordingly, histological sections from the same grey organ may show quite different pictures depending on the direction of the cut, and on the region from which the sample was taken.

A classification of the grey organs can be based on the class of possible different cuts which give indistinguishable histological preparations. Such a classification promises to be closely related to a fundamental classification of different types of transformation of patterns of excitation within the brain.

The considerable randomness which makes nerve nets different from crystals does not disturb us if we adopt the operational criterion of « distinguishable sections ». Thus « randomness » becomes equivalent to « no pattern », and the classification is based on characteristics which clearly emerge from the background of randomness. With this premise we may use the language of « symmetry » in order to define different types of nerve nets. The symmetry of a nerve net is defined by the group of the translations, rotations and reflections of the plane of the section which produce indistinguishable histological preparations.

a) *Full symmetry*: *random nets.* The statistics of the connections between two points of a random net is entirely defined by their distance. Hence any cut looks like any other cut and we may apply the concept of « full symmetry possessed by space itself » (WEYL). Many small grey organs of all brains are of this type. In the larger masses however we can always detect some degree of directional organization.

b) *Full symmetry of the plane*: *random plane.* The largest portion of the griseum in the human brain (progressively smaller as one descends the animal scale) is of this type. Cuts parallel to one direction (the « thickness ») are indistinguishable. In the direction of the thickness the statistics of the elements varies asymmetrically, *i.e.* there exists a non repetitive stratification (Fig. 11).

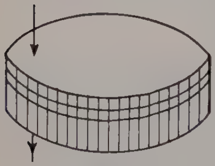


Fig. 11.

This type of griseum is called *cortex.* It appears generally in the form of large and relatively thin sheets whose extension may be (100÷200) times their thickness. Afferent and efferent fibers are usually related to two different levels of the thickness-stratification, which can therefore be recognized as the anatomical substrate of the transformation performed by the organ.

c) *Double translatory symmetry in the plane.* (Cortex with lattice structure). This type of symmetry is found in a portion of griseum which in all vertebrates is devoted to equilibrium and to the fine regulation of movement: the cerebellar cortex (Fig. 12). There are two entirely different sets of connections in two perpendicular directions of the plane, as well as a periodicity (about 50 μm in one direction and 200 μm in the other) which defines a clear lattice type symmetry. The direction perpendicular to the plane of the lattice is again asymmetrically organized. Even more striking than in the previous type is the sheet like expansion of the « cerebellar cortex », where the maximum is about (1 000÷2 000) times the thickness.

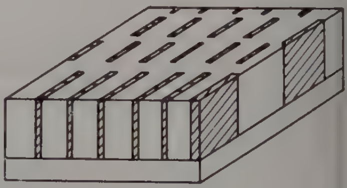


Fig. 12.

d) *Translatory symmetry in one direction* (the symmetry of a band ornament) characterizes the spinal cord, which is subdivided into identical segments, and where to a good physiological approximation the influence exerted by activity in one segment onto the next segment is alway the same.

e) *Translation and rotation* (through 90°) characterize the anatomical relation between two successive chiasmas and two successive cortices in the eye stalk of crustaceans.

f) Finally, it might be useful to mention some types of symmetry which do not occur in animal nervous systems. Thus the full symmetry of the plane, associated with translatory symmetry in the direction perpendicular to it characterizes a voltaic pile, but not any existing nerve net. Similarly, there are no organs with rotational symmetry, with the exception of the very loose and quite primitive nerve nets of animals whose body as a whole is of this morphological type (example: starfish, medusa).

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3° Trimestre

The Ascending Reticular System, the Regulation of the Sensory Inflow and the Problem of Visual Habituation.

G. Moruzzi

Istituto di Fisiologia dell'Università di Pisa
Centro di Neurofisiologia del Consiglio Nazionale delle Ricerche - Sezione di Pisa

My plan is to concentrate exclusively on one area of investigation, *viz.* the relationships between reticular formation of the brain stem and the transmission of visual messages to the cerebral cortex. There will be a further limit to the scope of the present review, since the problem of the interrelations between specific and diffuse projection systems will be approached from the angle of visual habituation. This drastic fence building is made necessary by time limits.

As an introduction, I should like to recall briefly the functional anatomy of the visual system.

1) The story begins with the *receptor-transducers*, *i.e.* with cells lying in the so-called sensory organs. They are endowed with the ability to transform physical or chemical stimuli into nerve messages. The rods and the cones of the retina are the visual receptors. They transform photic stimuli into visual messages. There are about 125 000 000 rods and cones in the human retina.

2) We are not interested here to know how the rods and the cones give rise to the nerve impulses constituting the visual messages. Let us state that the nerve impulses are conducted through two other types of retinal cells—the bipolar cells and the ganglion cells—to the optic nerve. There are about 1 250 000 fibers in each human nerve.

3) Each nerve impulse is a potential oscillation, whose size and shape is constant for a given nerve fiber and which is conducted at a constant speed in each nerve fiber. Hence any visual information which is conveyed to the brain may be characterized only by the patterns of spatial and temporal arrangement of the impulses coursing along the optic nerve.

4) Again we are not interested in the anatomical details of the cerebral projections of the visual fibers. It is enough to state that the impulses are relayed by the nerve cells of the lateral geniculate body to the visual area, which is located in the occipital lobes of the cerebral hemispheres. The visual system we have outlined so far is classified as a specific projection system, and the visual cortex is regarded as one ot the specific projection areas. The adjective « *specific* » means that all the structures considered above, and in particular the cortical projection area, subserve *exclusively* visual functions.

5) From the specific projection area the impulses may be transmitted to other areas of the cerebral cortex, whose function is to elaborate and to store the visual information.

I hasten to emphasize that this scheme has by no means been falsified by the new developments, that have occurred in neurophysiology during the las ten years. There is little doubt, however, that it must be completed with the results of more recent experiments (Rossi and Zanchetti, 1957).

First of all, we have recently learnt that the specific systems are not the unique channels through which the sensory organs may influence the brain. The visual impulses exert a generalized influence all over the cerebral cortex through an entirely different system, the ascending reticular system. This system takes its origin in the medulla and extends upward through the pons, and the midbrain. It receives impulses from the collaterals of all kinds of sensory pathways and persistently conveys the neural messages upward, where a diffuse system distributes them widely upon the cortex. One role of this ascending reticular system is arousal and the maintenance of the waking state. When this system is interrupted at midbrain levels the animal is deeply somnolent and it is only barely arousable momentaneously by intense stimuli. In this animal the visual impulses will still reach their specific projection areas in the cerebral cortex. However the animal is apparently unable to make visual discrimination or to recognize visual patterns.

Summing up, perceptual integration requires a close collaboration between the classical specific projection pathways and the ascending reticular system. Our first task will be to try to understand how such integration is made possible.

We are now coming to grips with a second concept, which we owe to the discoverer of the electroencephalography, Hans Berger. The importance of this discovery, which was made 30 years ago (1929), can hardly be over-emphasized, since it gave the demonstration that the neurons of the cerebral cortex are spontaneously active also when they are supposed to be at rest, such as during sleep or mental relaxation. The EEG, whether recorded from man or from animals, shows continuous fluctuating potentials at the surface of the head. Its most outstanding component is the alpha rhythm, which is characterized by a frequency of around 10 cycles per second. An important property of the alpha rhythm is represented by its disappearance when the eyes are opened. The co-ordinate pulsation of large group of neurons, which is responsible for the α waves, is supposed to be thrown out of synchrony in these experimental conditions. This disappearance of high voltage slow waves can be easily reproduced in all mammals whenever they are aroused by a startling sensory stimulation. It is called EEG arousal or arousal reaction and is observed throughout the dorsal extent of the cerebral cortex; it is probably mediated by the ascending reticular system (Moruzzi and Magoun, 1949).

It should be emphasized that all the new developments in neurophysiology would never have been possible without Berger's discovery. Let's consider again the visual system. From the retina up to the projection areas of the visual cortex the central nervous system was once thought to behave passively, *i.e.* one believed that the stimulus characteristics were simply transformed into spatial and temporal patterns of visual messages reaching the occipital lobes of the cerebral cortex. Berger's discovery and the findings which were prompted by it have definitely shown that the cerebral cortex is an active, dynamic mechanism whose responsiveness is continuously controlled by nervous structures lying in the reticular formation of the brain stem and in the midline nuclei of the thalamus. Indeed we have learnt that the retina itself is active in complete darkness and it is very likely that its activity and its responsiveness to the physical stimuli are steadily controlled by the central nervous system through efferent fibers coursing in the optic nerve. Thus sensory stimuli do not elicit, but simply modulate, an activity which would go on spontaneously in their absence.

The main conclusion to be drawn from all these considerations will be that a given photic stimulus may evoke quite different responses in the visual cortex according to the background activity of the visual cortex itself, the presence or the absence of other sensory stimulations, or any internal condition like emotions, bodily needs, and so on. This conclusion, after all, fits very nicely what we know from sheer introspection. Everybody knows that, for a given intensity ot physical stimulus, our sensation may be increased by attention or decreased by habituation. It is possible, however, to substantiate this assumption with objective, electrophysiological methods.

Whenever a volley of visual impulses impinges upon the specific projection area of the cerebral cortex a potential oscillation may be led from its surface. It is called « evoked potential » and is mainly related to the response of the neurons of the visual cortex to the afferent volley. Its size is usually regarded as a rough index of the number of the cortical units responding to the visual messages, *i.e.* of the intensity of the sensory response.

We shall be concerned here only with the changes of the photically evoked potentials during habituation. Habituation has been defined as « a process whereby certain sensory stimuli by repeated application lose significance to the individual ». Everyday experience indicates that this phenomenon is accompanied by a decreased awareness of the corresponding sensory stimulus.

Hernandez-Peon (1955) has recently quoted his unpublished experiments with Guzman-Flores and Alcaraz, showing that the evokes potential recorded in the subcortical visual relay, the lateral geniculate body, is decreased during habituation. He has reported, moreover, that the introduction of an auditory stimulus will abolish visual habituation (dishabituation). Quite recently Cavaggioni, Giannelli and Santibañez (1959) have confirmed these findings. They have observed, moreover, that the decline of the evoked response starts earlier in the visual cortex than in the corresponding lateral geniculate body; in other words, habituation occurs in the specific projection areas of the cerebral cortex when the subcortical relays are not yet habituated to the incoming volleys coursing along the optic nerve.

Thorpe (1950) has defined habituation as « an activity of the central nervous system whereby innate responses to certain relatively simple stimuli, expecially those of potential value as warning of danger, wane as the stimuli continue for a long period without unfavourable result ». The functional significance of the phenomenon is implicit in its definition, but we are yet some way from understanding its mechanisms.

Cavaggioni, Giannelli and Santibañez (1959) have shown that a cat falls asleep when it becomes habituated to a photic stimulus, at least if the animal is kept in a sound proof room and all other stimuli are avoided as far as possible. Does the animal fall asleep because the main stimulation is gradually losing its functional significance, as a consequence of habituation, or is visual habituation the result of the drowsiness of the animal?

The first hypothesis is disproved by an observation of Sharpless and Jasper (1956): cortical habituation may be absent, and indeed the primary cortical response may be increased, when the animal is beginning to fall asleep as a consequence of the monotonous repetitive stimulation.

Let us take into consideration the second hypothesis. It is seemingly supported by two important findings.

1) Dumont and Dell (1958) have shown that the potential evoked in the visual cortex by an electrical shock applied to the optic nerve may be greatly facilitated by stimulating the reticular formation, *i.e.* by experimental conditions reproducing the EEG arousal. An effect opposite in sign occurs during habituation, when EEG sleep patterns are present *throughout* the cerebral cortex.

2) Lindsley (1957) has recently reported a second group of findings, which also relate the evoked response to the ascending reticular system. In

the human subject two brief 10 microsecond flashes of light presented 150 ms apart are easily seen as two, and similarly at 100 ms separation. At 50 ms separation they are seen as one. In the cat and monkey the same two flashes of light at 150 and 100 ms separation will produce two distinct evoked potentials in the visual cortex, but at 50 ms separation only one such evoked potential pattern can be detected. If the reticular formation is stimulated electrically two distinct evoked potentials will be recorded even for intervals of 50 ms. Thereafter the response will revert to a single evoked potential as before reticular stimulation. Thus it is evident that stimulation of the ascending reticular system in some way facilitates the visual cortex, so that it can resolve two brief flashes. The conclusion might be drawn that just the opposite effect occurs when the animal becomes drowsy and that this reduced efficiency of the visual system during drowsiness is the cause of visual habituation.

The inadequacy of this explanation becomes apparent, however, on closer examination.

1) First of all if habituation were the result of a general decrease in the activity of the alerting or attention mechanisms, then it should be observed also for other sensory stimulations. There is little doubt, however, about the strict specificity of habituation not only to the modality (say the sound) but also to the quality (the pitch) of the repeated stimulus (SHARPLESS and JASPER, 1956).

2) The animal may fall asleep when habituation is not yet present in the sensory cortex (SHARPLESS and JASPER, 1955).

3) Although lacking after deep anaesthesia or in the coma produced by interruption of the brain stem reticular formation, habituation to an auditory stimulation can be observed in the cat during spontaneous sleep (HERNANDEZ-PEON, JOUVET and SCHERRER, 1957).

Hence the interruption of the sensory inflow which characterizes habituation is not causally related with generalized phenomena such as sleep or wakefulness, although it may be in some other way related to them.

There is little doubt that habituation arises in the higher parts of the central nervous system. If the blockade occurred in the retina itself or even in the lateral geniculate body, this would amount to a discontinuation of the habituating stimulus. Since it is well known that the habituated response recovers when the stimulus is discontinued, the conclusion should be drawn that sensory volleys still impinge upon the central nervous system when the evoked potential is absent or reduced in the visual cortex. Hence it would be perhaps wiser to state that during habituation the response to the visual volleys is different, rather than to say that it is abolished. The crux of the matter is to see what this difference is and why it comes to light as a consequence of a monotonous repetitive stimulation.

It would be dangerous to assume that the absence of an evoked potential means that no activity is elicited by sensory stimulation in a given nervous structure. It would be safer to state, say ,that during habituation the cortical neurons of the habituated cortex react in such a way when they are impinged upon by the visual volleys that an evoked potential can no longer be obtained. JUNG, von BAUMGARTEN and BAUMGARTNER (1952) and JUNG and BAUMGARTNER (1955) have clearly shown with their microelectrode studies that in the visual cortex there are several neurons whose response to visual messages is merely represented by an inhibition of their spontaneous discharge. This inhibition cannot be detected by leading from the cortex with macroelectrodes, as usually done in the electrophysiological studies on habituation.

Several considerations had already led to the conclusion that some kind of inhibition is responsible for habituation: possibly a phenomenon akin to Pavlov's internal inhibition. A peculiar type of inhibition, I hasten to add, since the same neurons will easily respond to the same sensory stimulation if other stimuli are applied simultaneously. This phenomenon, called deshabituation, hints that the inhibitory process which is responsible for habituation will fade away (or not appear) under the impact of other messages.

We are thus coming to grips with the core of our problem, a hard core and one which has not yet been solved. I am now going to present a few theoretical considerations, which may serve as a background for discussion and possibly for future experimental approach to our problem.

There would be no difficulty in thinking that sheer repetition of an identical stimulus would gradually increase the number of those visual cortical neurons which respond with an inhibition of their spontaneous discharge to the incoming volleys. This hypothesis may be wrong, but it is one anyway which can be approached experimentally. We hope to test it with microelectrode studies. If it is confirmed, the next step will be to investigate the mechanism of this gradual increase which is responsible for this type of inhibition.

It is usually stated that a servomechanism is an automatic regulatory device actuated by the difference or « error » between a desired reference input and the actual value of output. In this way a constant input is maintained, a phenomenon that is exemplified by the usual pupil reflex to light.

We are confronted here with an entirely different phenomenon, since it is the lack of an error which apparently evokes the habituation mechanism, whose tendency is actually to obliterate a constant input.

A reversal of the response of a single neuron—from excitation to inhibition—depending upon the background activity has been observed (VON BAUMGARTEN, MOLLICA and MORUZZI, 1954). The mechanism of the phenomenon is still unknown. The hypothesis might be advanced that any regular, monotonous repetition of an identical stimulus would bring about—by a phenomenon of temporal facilitation—an avalanching increase of activity in the inhibitory neurons, which are probably present both in the specific projection system and in the reticular formation. In the visual cortex a gradual decrease of the evoked response would occur, while at reticular levels this effect would be responsible for the appearance of sleep. Thus habituation of the cortical sensory response and sleep induced during habituation would not be causally related, although both phenomena would be the consequence of a basic property of the central nervous system operating at different anatomical sites. I hasten to say that inhibitory responses are elicited in several reticular neurones also by a single stimulus (von Baumgarten and Mollica, 1954). Our hypothesis would simply predict that these inhibitory effects would be strikingly increased by sheer repetition of the stimulation.

What we know about habituation might be explained quite well with this hypothesis.

1) Dishabituation elicited by other sensory stimulation would be due to inhibition of an inhibitory process, as Pavlov had already surmised. Conditioning obtained by a nociceptive stimulus would act in a similar manner.

2) General anaesthesia would depress more severely the inhibitory mechanisms, thus abolishing (or preventing) habituation.

3) The fact that the animals become habituated more rapidly on the second and third days of recording than on the first, hints that a state of habituation persists from one day to the next. This obviously relates habituation to learning, a similarity that had been already pointed out by Thorpe (1950).

BIBLIOGRAPHY

R. von Baumgarten and A. Mollica: *Pflügers Arch.*, **259**, 79 (1954).

R. von Baumgarten, A. Mollica and G. Moruzzi: *Pflügers Arch.*, **259**, 56 (1954).

H. Berger: *Arch. Psychiatr. Nervenkrank.*, **87**, 527 (1929).

A. Cavaggioni, G. Giannelli and G. Santibañez: *Arch. ital. Biol.* 1959 (in press).

S. Dumont and P. Dell: *Journ. Physiol.*, **50**, 261 (1958).

R. Hernandez-Peon: *Acta Neurol. Latinoamer.*, **1**, 256 (1957).

R. Hernandez-Peon, M. Jouvet and H. Scherrer: *Acta Neurol. Latinoamer.*, **3**, 144 (1957).

R. R. Jung, R. von Baumgarten and G. Baumgartner: *Arch. f. Psych. u. Ztschr. Neurol.*, **189**, 521 (1957).

R. Jung and G. Baumgartner: *Pflügers Arch.*, **261**, 434 (1955).

D. B. Lindsley: *Nebraska Symposium on Motivation* (Lincoln, Neb., 1957).

G. Moruzzi and H. W. Maugoun: *EEG. Clin. Neurophysiol.*, **1**, 455 (1949).

G. F. Rossi and A. Zanchetti: *Arch. Ital. Biol.*, **95**, 199 (1957).

S. Sharpless and H. H. Jasper: *Brain*, **79**, 65 (1956).

W. H. Thorpe: *Symp. Soc. Exp. Biol.*, **4**, 387 (1950).

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3º Trimestre

Men and Information: a Psychologist's View.

E. B. Newman

Psychology Department Harvard University - Cambridge, U.S.A.

Introduction.

In the context of a Course on information theory, a psychologist feels strongly tempted to start his speech with a series of caveats. I will say only this, that in spite of the assurances that you have been given it is still not true that a man can be considered *a priori* as anything predictable. I would also say that all the assumptions made by mathematicians and engineers about men were certainly false, so that if I contradict anything said by anyone else up to this point, I am asserting categorically that they are wrong and I am right.

It is perhaps almost a commonplace that a man can be regarded as a communication system. He receives sensory impressions (information) from his environment (a source). This information is transformed (recoded) initially in the sense organs and subsequently in successive centers of the nervous system. It is transmitted over nerve trunks from one station to another and amplified in the course of its transmission. It is stored both for short times necessary to effect sequential recoding and for long times in something that, remarkably enough, is called memory. Finally, the information is retransformed into environmental events that contain some fraction of the information that entered the sensory end of the channel. The system is noisy as any real system is bound to be, but it is amazingly economical, highly portable, easily programmed and has been proved out in service in essentially the present model for at least 5000 years for which we have records. Quite a machine!

At the same time, many people have had the reverse idea, that nature, including machines, often imitates the behavior of men. Primitive man ascribed human motives to the wind and storms, to the mountains and to nature about him. He saw devils and spirits in the inanimate world. The recent fascination with robots and « Giant Brains » has about it some of the same uncritical bemusement with physical systems. The building of man-like machines will not occupy me here because I feel that sweeping analogies are cheap and contribute rather little to understanding human behavior.

Before discussing specific experimental results, it should be pointed out that any given physical unit may be regarded as any given part of a communication system depending entirely upon what our purpose is. Thus a man is clearly a source of a message or a destination when the engineer is talking about a communication channel such as a telephone. But he may equally well be regarded as the channel between some physical source and behavioral end result. Or again his behavior may be the source for experimental and statistical procedures leading to some inference.

At the present time it is most difficult to say very much about a man as a source. We know something about the statistical properties of some of the signals that he emits, such as the sequence of sounds of speech. Recently there have been hints that more can be done with this area, generally in the direction of statistical models, of which a MARKOV process is a simple example. I am referring to recent work of NOAM CHOMSKY and GEORGE MILLER. Unfortunately, I am not competent to discuss these developments.

It would also be possible to describe a man as the receiver of signals and, in a sense, someone's ability to receive information is almost always presumed as the goal of any system. Machines do not run for themselves but they are created to serve men. Once more this is a problem beyond our present abilities and I shall mention it later only in an oblique way.

This brings me to the third possibility, and that is to regard man as a channel. Such was the general description with which I started. What I am going to do is to discuss the input end, that is, sensory and perceptual processes which encode the input. Then I shall turn briefly to the output end and return finally to the central part of the process.

1. – Information in single perceptual displays.

One of the most obvious questions to ask in terms of information theory is what is the capacity of our senses, the eye or ear, to transmit information? How does it compare with the channel-capacity of a telephone or television circuit? Before trying to answer this question let me make three points to clear away some ambiguities.

1·1. Most discussions of channel-capacity, and particularly the relation of channel capacity to band width, are based on signals considered in real time. On the other hand, when we are dealing with speech, with symbolic and logical operations, and with other human activities, the natural scale is in terms of events. Speech consists of a sequence of phonemes. The information conveyed is invariant with number of phonemes rather than with real time which is a less meaningful variable. The two types of scale are not unrelated, of course, but in our first approach to many problems we shall use the scale that is most convenient, and shall leave for the future the reconciliation of data from the other.

1˙2. In discussing the question of channel-capacity we are indulging in an illegal activity. Specifically, we are asking certain empirical questions rather than formal questions. As it was originally introduced by Shannon, the capacity of a channel, C, is a quite formal concept. The empirical questions on the other hand, are interesting ones and I believe they are scientifically fruitful. I hope that it will remain perfectly clear that my use of the term channel-capacity does not carry with it any formal implications.

1˙3. In dealing with an informational analysis of human behavior, there is one very serious limitation that must be studiously ignored. This is the problem of sampling. Let me take speech as an example. For purposes of analysis we assert that certain sequences of words have a certain very small yet real probability of occurring. Actually, in any practical sample they never occur. The probabilities are too small. And it would take more than a man's lifetime for him to utter enough sentences to constitute a good sample. On the other hand, some utterances do occur, and the fact that they occur may well be statistically highly improbable. To add the *coup de grace* to our other troubles, we note that people are not highly stable systems but they change as a consequence of experience. No man is ever twice just the same person. Thus we conclude that in fact no statistical model can ever be proven correct for all the phenomena with which it is supposed to deal. It is just one of these lovely fictions, a bit of the *Als-Ob* that makes scientific life worth living.

a) One way of estimating the capacity of the human senses is to build upon the observation that the sense organs and the nervous system are discrete in certain important ways. First of all, the peripheral sense organs and the nerve trunks that serve them are made up of discrete cells capable at times of independent action. Furthermore, the action of most nerve cells is all-or-none; a discharge once started is propagated at maximum amplitude, and such a discharge is followed by a refractory period before another discharge can be initiated. Thus, it is easy to suggest that the nervous system is like a multiple-wire cable over which pulses are transmitted. Each pulse on each wire represents a point in a probability-space, and by the same token so does each discharge over each fiber. From these observations one could calculate a rate of transmitting information. I am persuaded that this picture is too simple but I shall not go into that now.

Another well-known stepwise phenomenon is the existence of a threshold in sensory responses. One such threshold is the minimum strength of a stimulus that can be reported. Below the threshold there is no reportable effect; above the threshold, an observer sees a light or hears a sound. There is also a threshold for *differences* in any discriminable aspect of stimulation. Given a sound of any loudness, a second tone will be heard as equally loud until the intensity has been increased more than a threshold amount. These two thesholds suggest a kind of psychological scale with discrete steps for each sensory attribute.

A very little consideration makes clear that this view is also far too simple. For one thing, the existence of the absolute or stimulus threshold says nothing about the magnitude of supraliminal sensations, and therefore nothing about how much information is contained in an experienced magnitude. It sets a kind of a boundary, and that is all. Moreover, there is an ancient paradox connected with the differential threshold. Suppose that you can just distinguish between 100 grams of weight and 110 grams. What will happen if you now take 105 grams as your standard? Will the sudden step still come at 110 grams? The experimental fact is that the noticeably heavier weight will be 115. Quite evidently there is a continuous variable that underlies the saltatory effect that we call the threshold. Another form of the paradox is that 100 and 106 will be judged equal. So will 106 and 112, 112 and 118, and yet 100 is very different from 118. In logical terms, the equation is *not* transitive, for 100 and 124 are easily distinguished. In our present context we are interested in the reverse argument: it is not proper to argue from the fact of a threshold to the concept of a stepwise scale. The facts suggest that our variables are in some cases continuous (in computer terms, analog) transforms and that the threshold is a particular function (digital) superimposed on the underlying scale.

In passing it should be mentioned that both the discreteness of nervous response and the threshold have been used as a basis for calculating the information received by the ear or by the eye. These estimates turn out to be very large; I shall not repeat them here because they are pretty obviously wrong.

b) As an alternative procedure for measuring the capacity of the sensory channel we turn to more direct measurement of transmitted information. The procedure is straightforward. A set of stimuli, such as notes on the piano, are presented to a subject who identifies the notes presented in accordance with a pre-arranged code. As the size of the set of stimuli is increased a point is reached where there are a significant number of confusions. From the resulting input-output matrix (or confusion matrix) it is a straightforward matter to compute transmitted information.

Now for some experimental results. Pollack (1952) has reported results for pitch. From 2 to 14 different pitches were presented with results shown in Fig. 1. Up to 4 pitches were identified perfectly; beyond that number, errors increased rapidly and the curve becomes asymptotic to a limit at about 2.3 bits per stimulus. Changes were made in the range of frequencies employed, in the distribution of stimuli within the range, and in the loudness of the tones. These, and doubtless many other possible variations, do not increase significantly the information transmitted.

Both POLLACK (1953), POLLACK and FICKS (1954), and GARNER (1953) have done the same experiments with loudness. Tones of various intensities are presented and the listener has to label them accurately. Fig. 2 shows GARNER's results after partialling out disturbing intraserial effects. Once more, the curve rises with a slope of 1.0 to just over 2.0 bits per stimulus and then levels out with 2.3 bits as a limit.

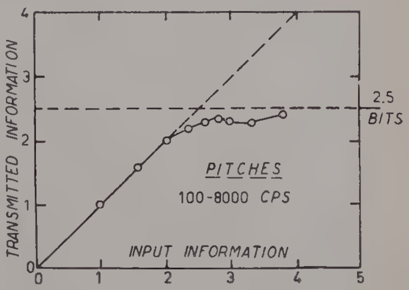


Fig. 1. – Information trasmitted by presenting one of a set of tones (POLLAK, 1952).

The saltiness of salt was tested in the same way by BEEBE-CENTER, ROGERS and O'CONNELL (1955) with results shown in Fig. 3. The tongue is less acute, according to their results, than is the ear. The maximum lies at about 1.9 bits per stimulus representing an accurate discrimination of about four levels of saltiness. But while the decrease from 2.3 to 1.9 bits is significant, it is trivial in contrast with the broad fact that information is almost invariant when stimulus magnitudes are chosen from widely different ranges. GARNER's tones covered a range of 95 db while the salt solutions varied in concentration by a ratio of 1 to 100. (We might call this 20 db).

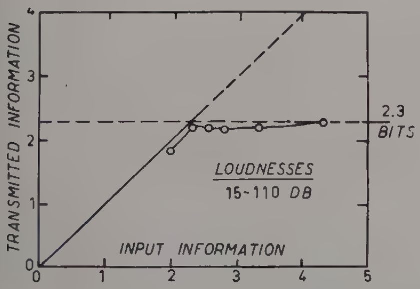


Fig. 2. – Information trasmitted by presenting tones of varying intensity (GARNER, 1953).

The eye does slightly better, in some respects. HAKE and GARNER (1951) report results from an experiment in which subjects estimated the position of a pointer between two marks on a scale. The stimulus positions that were utilized divided the interval into 5 parts, into 10 parts, into 20 parts and into 50 parts. The results are shown in Fig. 4, in which the open circles indicate results when the subject knew how many positions were being employed, and the filled circles judgments without knowledge of the stimulus positions. The asymptotic value of 3.25 bits per presentation is about one bit higher than the values for pitch and loudness. It is only fair to point out that in this case there is one important difference from the previous experiments. This is the simultaneous presence of anchoring stimuli at both ends of the scale. POLLACK found that a single anchor in the case of pitch judgments added roughly 0.5 bits. There is the further fact

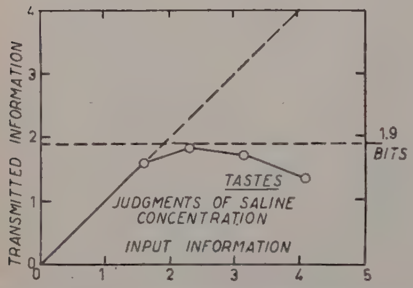


Fig. 3. – Information transmitted by salt solutions of varying concentration (BEEBE-CENTER, ROGERS and O'CONNELL, 1955).

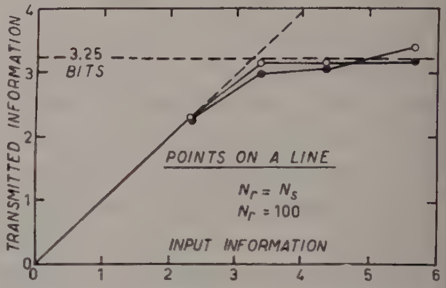


Fig. 4. – Information transmitted by positions of a pointer along a bounded line (HAKE and GARNER, 1951).

that one of the stimulus positions in HAKE and GARNER's experiment was directly over the scale marks, allowing for no uncertainty in the report. These two factors may account for as much as 0.7 bits. Subtracting this, the estimate is once more in the same range.

The single aspect of size did less well in an experiment of ERIKSEN and HAKE (1955) using squares of different sizes. Since the shape was constant this presumably is equal to a judgment of linear extent. Their highest values were 2.2 bits, comparable with the corrected estimate of 2.5 for a point on a line. POLLACK (see MILLER, 1956) has reported values of 2.6 and 2.7 for area, and 2.6 to 3.0 for length of line.

While the experimental work is far from being exhaustive, the sampling of stimulus dimensions has been reasonably broad. The invariant relation that emerges is rather striking for a psychological function. Whenever information is conveyed by systematic variation of a single psychological attribute there is a limit to the number of useful steps that may be employed, and this limit lies between four and eight steps, *i.e.*, two to three bits of information, over a wide range of conditions. Once the steps are clearly above the differential threshold, neither the range nor location of the stimuli seem to be of great consequence, nor is the time allowed for judgment, nor the form of the report. The only added variable of significance is the use of explicit standards, and in this instance it appears that not the so-called absolute but rather a differential judgement is added in. In some very basic aspects the use of information from absolute judgements is a very different process from the discrimination of a difference.

c) In what has been said up to this point about channel capacity there has been an important basic limitation, namely, that we are sending a single message by the choice of a value from a single dimension. The choice turns out to be from a rather small set of alternatives, let us say, five or six. But a man is not limited in this way. Objects and events vary in many more ways than a single attribute and presumably a response can be made at one and the same time to several of these attributes. Now the question is, are these attributes independent of each other, or, does the information in each add to the information in the others? The answer is that additivity is sharply limited.

Once more I should like to interpolate for this audience the remark that this is an old question in psychology. It is the problem of attention. It has long been known by psychologists that the set (attitude, Aufgabe, Einstellung, determining tendency) can be a major determinant of the report that an observer gives, and, particularly, that a person set to observe one aspect of an object exposed for a brief moment will be less able to report some other aspect that might be clear if the set were changed. This marked selectivity of the human and, we believe, animal observer must be familiar to all of you. What is disturbing is that this filtering, this pre-tuning of the receiver, changes with great rapidity and we mistake the sum of a series of successive impressions for a single simultaneous one.

Please do not think that the problem is solved by these remarks. It is in fact only made more difficult. While it may be true that the demands made on the input channel are found to be less severe by virtue of the selectivity of attending to one thing at a time, there are posed for the organism two other problems that are quite as difficult. These are, first, the problem of what determines the set, and the rapid changes in set, in normal observing, and, second, how is all this successive information stored and organized so that a man can act as if he had a very large channel capacity at any one moment.

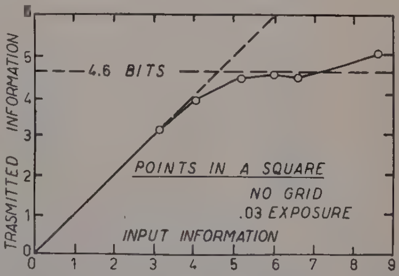


Fig. 5. – Information in position of a point in a square grid (KLEMMER and FRICK, 1953).

Having argued against the complete additivity of various dimensions, let us look at some experimental results. What does happen to channel capacity as the number of dimensions in the stimulus is increased? In a study by KLEMMER and FRICK (1953) the issue is put quite clearly. They used a pattern with a dot, shown on a screen, and asked subjects to mark down on a paper with squares where the dot was located. Actual grid lines on the target and on the record sheet made no difference in the results. The fact that the position of the dot varied along two coordinates made quite a big difference. The results are shown in Fig. 5. As the number of possible positions on the grid increased, the information in the source increased, and the information transmitted went right along with it until a fairly sharp limit was reached, at about 4.5 bits. This is to be compared with 2.5 bits for a single dimension.

Klemmer and Frick carry the case a step farther. Suppose there are two dots in the square, then are four coordinates. And indeed they find that the limit rises again. Three dots can be thought of as six coordinates, and so on. The actual tests were limited to four dots in a 3 × 3 matrix because the total set becomes so large as to make any exhaustive sampling impossible. But the point that is clear is that the losses in transmission are quite small. In their most complex case, from 1 to 4 dots in a 3 × 3 grid, they were able to recover 7.8 out of 8.0 bits of information. Klemmer (1957) has recently extended this study to a small sample of the possible patterns on a 4 × 5 matrix. For his best subject he reports values between 11 bits and 13 bits. There are very large differences among individuals in a short test and it seems probable that the training of the subject in reporting becomes very important. Nevertheless, we are obviously moving to a fairly high level compared with one-dimensional tests.

I should like to comment briefly at this point on the ambiguity of the term attribute, or aspect, or coordinate. In the Klemmer and Frick study it is possible to think of four dots each having two coordinates and eachco ordinate having three possible values. The stimulus information is then $4 \times 2 \times 1.6 = 12.8$ bits. But it becomes clear that this estimate is too high. Certain instances are excluded because they are identical. Obiously, it is more meaningful to map the whole thing into a nine-dimensional space — one for each of the cells — and think of each dimension having possible values of 0 and 1. Thus there are 2^9 possible patterns — and nine bits of information. I recite this elementary instance because it is typical of a very large class of problems. The issue is what particular language is useful in describing the situation to which a person responds. Depending upon one's choice in this matter, man can look quite ineffective or quite effective. Or perhaps it is the person asking the question who looks wise or stupid.

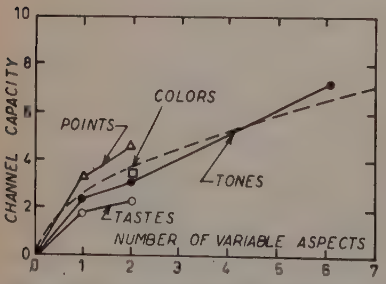


Fig. 6. – Summary of information transmitted by multidimensional displays (Miller 1956).

It is not worth the time to describe other multi-dimensional studies. Several have been done, and none of them are completely satisfactory. One trouble is that any systematic testing of displays with, let me say, 30 bits of information is a rough job. The thirtieth power of two may be a small number to people who are dealing with orders of infinity, but when an experimenter tries to sample exhaustively such a set, 2^{30} looks enormous.

A few results are summarized in Fig. 6. The point to be made here is that additional dimensions do add more information, but the return falls off as the number of dimensions goes up.

d) There are several lines of indirect evidence to suggest that there is an upper limit for the channel capacity, no matter how many dimensions are chosen, at 15 to 20 bits per presentation. There is, for example, a good deal of rather old work on the so-called span of apprehension. A set of letters or numbers is exposed in a tachitoscope foı a brief period of time — perhaps 1/50 s. The subject reports what he has seen. Four to five randomly selected letters represent an average man's limit, which adds up to perhaps 20 bits. It is also known that the span increases as the choice of material is limited to words. Clearly the number of items goes up as information per item comes down.

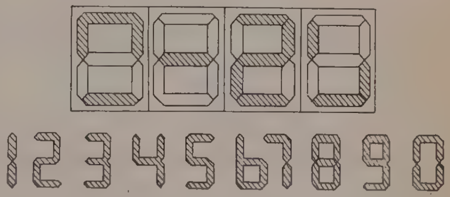


Fig. 7. – Top: appearance of the stimulus field of four light boxes, each containing seven individually lighted line segments. In the typical random patterns of lights shown, the black area represents the lighted lines. Bottom: patterns used to represent arabic numerals.

Another line of evidence can be derived from eye-movement studies of reading. A good reader makes as few fixations as he can without retracing more than occasionally. In one recent study using college students, an average of eight fixations for a line of 60 letters were found (CARMICHAEL and DEARBORN). This works out, at 2.0 bits per letter in meaningful text, to 15 bits of information per fixation.

I would like to mention finally a quite recent study of KLEMMER and LOFTUS (1958). They used a stimulus display shown at the top of Fig. 7. In their main series of tests they used a random selection among the 128 possible patterns at each of the four positions. These included patterns that might be interpreted as numerals or letters. The forms of the numerals is shown in the bottom of Fig. 7. One obvious result of the experiment was that the subjects had trouble in writing down what they saw. Usually 5 or 6 s were required and much of the figures had been forgotten by the time the subject got to it. KLEMMER and LOFTUS resorted to an important technique, namely telling the subject after he had seen the set of figures what part to report. This was done, for example, by a small red light under the figure. The results are given in Fig. 8. They also found very substantial practice effects and some preference for continuous or closed figures. Under the best conditions — partial report with a prompt cue and some three weeks of daily practice — the best subject got up to 26 bits, the average subject up to 22 bits of information perceived. The best measure of transmitted information, that is, instances actually reported totally, was about 18 bits. Interesting enough numerals and letters did *not* show any superiority over similarly shaped figures. Familiarity helps, if at all, in remembering the material long enough to permit reproduction. But this is another question.

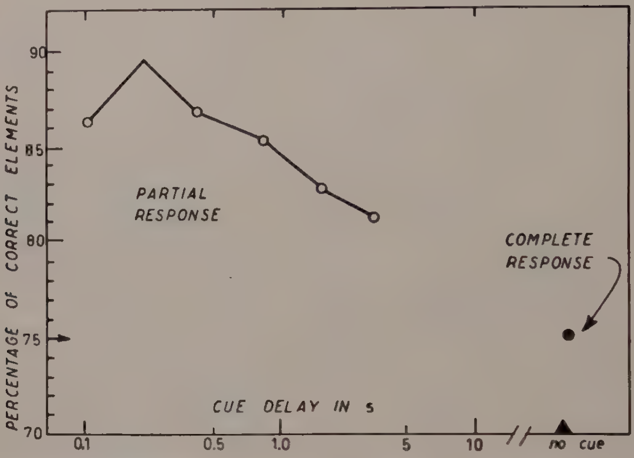


Fig. 8. – Percentage of elements correct as a function of the time between stimulus and poststimulus cue. Performance on test requiring complete response is shown for comparison. Data average of ten *S* s.

To recapitulate very briefly, we found that the information conveyed by an item presented along a single dimension was quite sharply limited to about 2.5 bits and that this amount was invariant over a rather wide range of conditions, sufficient to suggest that we are dealing here with a basic if ill-defined psychological mechanism. We have also shown that the information increases, although not quite proportionately, as the number of dimensions is increased. Just what constitutes a dimension or coordinate is left unclear. But even as the number of dimesions is increased there is still a limit on the perceptual system. This limit is probably between 15 and 25 bits per presentation.

2. - Perceptual capacity in real time.

Let me turn briefly from this kind of an analysis in terms of short but static presentations to the problem in real time. Can we translate from estimates of information per event, or per presentation, to a continuous flow. The most immediate observation to be made is that the limit on the « channel » may occur at quite different points. Thus Klemmer and Loftus were able to get into their subjects some 25 bits with an exposure of 1/50 s. Actually the time might have been 1/10 000 s with an extremely bright light. Obviously such very short times cannot be converted into a rate such that 50 or 10 000 fields could be usefully presented in a second. For one thing, response is seriously limited. Even to get evidence of 25 bits, a procedure had to be adopted that called for only a 6-bit output. Furthermore, the sensory mechanism itself is limited. We know from studies of flicker fusion that randomly selected fields at 50 per s would fuse into a uniform grey and nothing would be seen. Once more it is important that we avoid too hasty inferences and turn rather to a direct measure of a rate in real time.

I do not want to stop here to review studies made on what I might call the « straight-through » rates, that is, rates in which the input information is reproduced in substantially the same form in the output. These include typing, reading aloud, sending and receiving telegraph, piano playing, taking dictation in short-hand, etc. Let me just say that all of these rates may well be limited on the motor side as well as the percentual side, as has been pointed out by QUASTLER. For the present I would like to limit my consideration to problems of sensory or percentual inputs in real time.

Of the various kinds of evidence, the speed of silent reading gives us the highest rates on which we have any reasonable evidence. Successive fixations of the eyes occur at an average rate of about four per second. If 15 bits are perceived in one fixation, then we presumably receive information at a rate of 60 bits/s. Quite obviously most of this information is discarded almost in the moment it is read, for no one has the ability to recall or to recode information at this rate.

The other empirical approach to a rate of discrimination in real time is through auditory, and particularly verbal, material. The estimates of information are less clear cut than in the case of reading because there is somewhat less agreement on the proper segmentation of a flow of speech sounds. For example, a fine phonetic analysis carries the implication that the listener hears much finer discriminations than he does if a loose phonetic analysis is made. In addition, there are clearly nuances of meaning and of emotional expression that are carried in the auditory pattern but are missing from the written material.

The problem of segmentation is closely related to the problem of the choice of a level of analysis. Should it be in phonemes, or morphemes, or words, or sentences? These are areas into which I do not care to venture, especially as these matters are discussed elsewhere in this Course. What is important for the present purpose is to assert that as yet there are no radical differences in the estimates of information rate depending upon the level of analysis employed — nor is there any great difference between auditory and visual.

Let me put the matter in this way. Presumably a trained listener can understand speech sounds somewhat faster than the maximum rate at which they can be spoken (an exception might be a speaker highly trained with redundant material and a listener able to discriminate but not in possession of the full code). The best guess is certainly that the ear can hear between 30 and 50 bits per second.

To sum up in this question of our discriminative capacity in real time, I should like to make the following points:

1) In any given example there may exist a limiting rate either in the perceptual input, in the processing and recoding of the information, or in the motor output. Presumably the lowest of these establishes the overall rate.

2) Except for certain straight-through rates the commonest limitation is on storage and output. No person at the end of a minute is ready to do any one of a billion (10^9) equally probable different things. Yet his senses may have provided him with enough information to make such a choice.

3) Serious as the mis-match between input and output may be within the organism, it is even greater between the outside world and the organism. The extreme example is, of course, a television channel. I shall want to say more later about this discrepancy.

4) The actual use of a channel in every day life is usually at a fraction of its maximum rate. For example, aircraft pilots receive information at a rate of perhaps 2 bits/s over their radio telephone channel. The principal advantage of these low rates is the protection they afford against noise and error.

3. – Channel capacity in motor output.

Somewhat further removed from information theory is the study of man's output, particularly in terms of the work that he does. These matters have been of interest both to psychologists and physiologists for many years. Thus we know a good deal about how strong people are, about the force and energy with which various movements can be carried out. There are detailed studies of muscular fatigue. There are studies of the precision of movements, various kinds of aiming tasks, and of rates of movement, such as tapping and various measures of reaction time. But almost none of these studies touch on problems that can be put into informational terms.

The simple fact is, of course, that the behavioral repertoire of most animals is amazingly limited so that the selection among various movements that an animal can make carries rather little information. Man is fairly far out on one end of the distribution in two important respects. First, he appears to have somewhat greater differentiation of his vocal apparatus than most animals (although perhaps not some birds) and he has independence in the movement of his thumbs and to a less extent of the other fingers, so that he has a fairly high degree of manipulative skill in his hands. But these are differences in degree and at best they are not great.

What becomes obvious from an inspection of the gross facts is that the tremendous difference in informational output between a cow or a hen on one side and a man on the other depends relatively less upon the size of the set of possible movements but rather on timing and the sequential patterning of behavior. Let me take two simple examples to make my point. A telegraph operator makes only one or two basic types of movements involving perhaps a dozen muscles. It is no more complex, in fact probably less complex, than a hen pecking a grain. But the latter movement is utterly stereotyped and the former transmits the contents of a newspaper. Again, studies have shown that the corrective movements made to the « stick » in flying an airplane, or to the wheel of a car, are not finelly graded, exact movements. The precision required in flying a plane from New York to Paris is achieved by making a series of 2-bit responses at the right time and in the right sequence.

After this rather long introduction, let me describe briefly a few examples of studies that have used informational estimates of behavior. The first have to do with reaction time. This is the time required for a subject to react to any specified signal. Characteristically he is given detailed instructions before the signal so that he is « set » in a particular way. Most often reaction time is measured in a series of single tests, but it is possible to arrange the tests in such a way that each response triggers the next signal and one measures a serial reaction time. The results are not greatly different.

Reaction time has been known to be fastest to a single signal. It is substantially slower if the subject has to press one of two keys, or to respond or not respond, depending on the signal presented. Such choice or conditioned reactions can be among several alternatives. Starting from this point HICK (1952) showed that it was possible to account for reaction times on the assumption that information was being transmitted at a constant rate. He had to make one strong assumption that has not been readily accepted. This is the assumption that the simple reaction time represents a 1-bit, not a zero-bit choice, and that generally the rate is calculated as $\log(n+1)$ were n is the number of actual alternatives in the disjunctive reaction. HICK found the fit of his data to this formulation rather good. Other experimenters have not agreed. The alternative view is to suppose that there is a minimum transmission time that has nothing to do with the processing of information, and that there is added to this a processing time that is proportional to the information required to make the choice. Fig. 9 taken from a study by HYMAN (1953) strongly supports this second view. HYMAN used a variety of techniques for changing the « information » in the stimulus input and finds a relatively invariant relationship of information to reaction time.

Not all such studies have been so successful. Quastler and Brabb (1956) measured reaction time on a typewriter using trained typists. For the full alphabet the reaction time was .53 s. For an 8-letter alphabet it dropped only to .52 s: for four letters to .49 s; and for a single known letter to .21 s. This is very far from either Hick's or Hyman's result. As Quastler and Brabb suggest, the short-term instruction is probably not effective in reducing the complexity of the task in the face of a history of many hours of practice with the full alphabet.

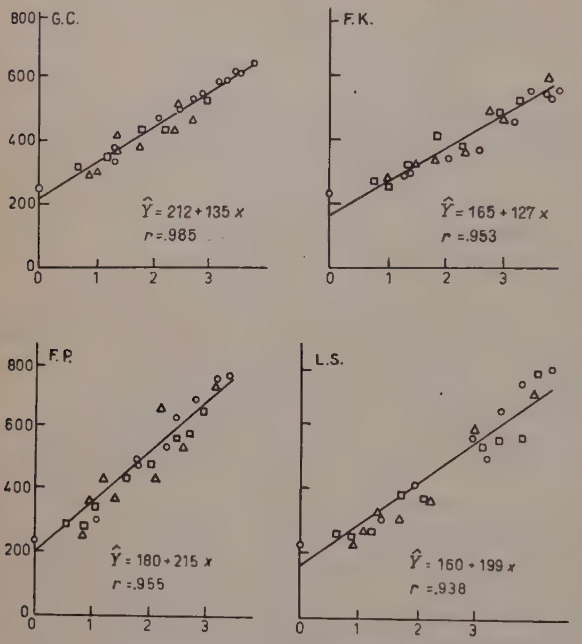


Fig. 9. – Reaction time (expressed in .001 s) as a function of stimulus information (expressed in bits) when amount of information is varied in three different ways (○ Exp. I, ◻ Exp. II, △ Exp. III). Data are for *S* s G.C., F.K., F.P. and L.S., respectively.

Klemmer and Muller (1953) report results that look much like those of Quastler and Brabb. Their task was not a simple disjunctive reaction but rather was one that required that a pattern of fingers be pressed simultaneously. Moreover their tests, with one exception, were paced. This makes it impossible to compare the results directly. Their date suggest, however that with only one stimulus (0.0 bits) the optimum spacing is about .25 s while for the 2.0, 3.0, 4.0 and 5.0 bit task, the reaction time was about .40 s. The slope of Hick's functions is missing. It is not immediately evident just why the two, so nearly parallel experiments give different results. Perhaps the use of five keys simultaneously instead of the single disjunctive reaction is a parallel to the dimensionality of a display. That is, Klemmer and Muller have added not only information but also more dimensions to their task.

High output rates are found only in highly-skilled tasks. Presumably they give us a pretty good estimate of top human capacity. I would mention these very briefly. One is the rate of talking — or more properly — of reading. With optimum vocabulary, rates of 24 bits per second have been obtained. With limited or highly redundant vocabularies the informational rates fall off because of mechanical limitations in the production of sounds. Much the same thing can be said about typing and piano playing. There is a psychological limit on the speed with which the fingers can be moved. With a low informational output per movement, the limit is the mechanical one. With more complex material and more complex coordinations, information per movement goes up. Total information also goes up until the subject begins to make errors unless he slows down. Somewhere in this region there is characteristically a plateau, with a rate of movement not far below the maximum possible peak information in the neighborhood of 25 bits per second, and the information per element of the motor performance up to perhaps 5 or 6 bits, but certainly not very high.

It is still an open question whether performances of this kind could be achieved with quite different motor systems if the subject had years of practice. There is a fair chance that in the course of social evolution we have evolved techniques of communication that make optimal use of the human organism. But there are just enough places where we are sure that social evolution is not maximally adaptive so that we should be reasonably skeptical of any very dogmatic statements that these performances are the very best that any man will ever do.

4. – Memory.

From all that has been said up to this point it must be clear that if one looks at input and output rates alone, a man does not turn in a very impressive performance. Surely there are capacities of the human organism that are not topped in these terms alone. Where is the trouble? Do we overestimate ourselves? Are we really just a somewhat overworked 25-bit channel? Or are the facts that I have supplied you grossly in error?

First of all I would say that I am strongly convinced the facts are correct. Admittedly this is a judgment but by now enough people have been over this ground to make the main points pretty clear. Well then, is a man such a poor machine? No, I think not for two reasons. First of all he is not one but a very large set of 25-bit channels. As I have remarked earlier, the rate at which a man reprograms himself appears to be high. Let me just repeat that in many ways this problem of set is almost the key problem with which we have to deal. Let me remark that I believe I am pointing to the same problem that W. Rosenblith is going to describe as a « state » variable. The second reason man does so well is because he makes use of a substantial memory. It is to this problem I should like to turn.

One very simple requirement of the human system for storage is a consequence of the redundancy of our inputs and outputs. The signals furnished to us by nature are highly redundant and the signals that we generate are also redundant. Much of this redundancy is of almost no consequence at all. Thus it is trivial that a visual signal has any particularly coherence within a time period of less than 1/100 s. The photoreceptor processes are such that they respond to an average of the radiant energy over a period at least this long. There are many other instances of the same kind in which the possible inputs simply fall outside the range to which a person could possibly respond. But there can equally well be coherence within the range of possible response but where the responder makes no use of the redundancy. For one thing, there is often no stress. Most of us walk about the streets or our homes making use of visual cues under circumstances where we could probably do almost as well if we were blind. At least in terms of the tests that we perform, there is little or no evidence that we make use of the strongly habitual character of many acts.

Speech communication is perhaps one of the bright spots. Certainly many of us try to listen under stress. Conversation at a cocktail party or on a subway train is marginal. I am sure you would all testify to that. And were it not that cocktail party conversation with a pretty girl requires only about a two-bit answer, it would be totally impossible. More seriously, the evidence is that we depend very critically upon the redundancy of speech and if we are to make use of this redundancy we need a good storage system. G. Miller (1956) has remarked, and I agree most heartily, that it is the immediate context that is of paramount inportance. In the case of speech most of the constraints are contained in the 10 to 15 adjacent letters. But this means that there must be a running storage of perhaps 3 to 6 words held in somewhat raw form that can be reinterpreted on the basis of the total pattern.

Actually there is rather little detailed evidence on this kind of short-term memory and just how it functions. Two findings have a possible bearing, however. These are the immediate memory span and the eye-hand span. To start with let us look at the immediate memory span. If a series of numerals are spoken, then a subject can repeat immediately a span of from seven to ten digits. This test is a very familiar one because it has been used for years as part of the Binet tests of intelligence. The span is highly correlated with the age of children.

What happens when we use letters or words instead of digits? Is the informational content of the stored material in any way constant? It is found that the span falls off slightly as the material is drawn from a larger set. In an experiment by Hayes (1952; see also Miller 1956) the span decreases from about nine digits to about five words from a 1000-word vocabulary. These results are rather far from being invariant when converted into informational terms. Nine digits are 28 bits and five words represent at least 50 bits. Other results such as Miller and Selfridge (1950) on the one end and Smith (see Miller 1956) on the other end, suggest that this function is rather flat, but I am inclined to be quite skeptical of both results. For one thing, the method used by SELFRIDGE to construct material leaves the informational value of her words in doubt. At the same time, the fact that SMITH (and HAYES) used limited sets of numbers is relatively artificial. I am doubtful that telling a subject he now is dealing with 1 and 0 is sufficient to change his set from the usual set of decimal numbers.

There is the further difficulty that this type of experiment is in a way self-defeating. To the extent that numbers contain less information, more of them must be retained and this requires a longer time. But if the process under investigation has some progressive decay function then the longer list will be less well retained. There is the very considerable further problem that the span is measured in isolation and the process of using the context in understanding information through a noisy channel is a relatively continuous one.

And yet — in spite of all these difficulties and limitations — it is still true that the order of magnitude of the context necessary in order to utilize the redundancy of language is of the same order of magnitude as the immediate memory span.

The other line of evidence comes from studies of the eye-hand span. Now there is in any real informational processing link a certain delay. This can be thought of as a straight transmission and processing delay, and it is very analogous to the reaction time measured in the isolated disjunctive reaction experiment. Let us consider what happens as a person copies a telegram or a radio message or typewrites from a copy at which he is looking. Many of the movements made, such as pressing a key, take far less time than the 0.2 to 0.3 s required for a reaction. Obviously the second action must start before the first is completed, and there easily develops a lag in the motor process so that the eye is reading or the ear hearing well ahead of what is being written. The same thing is true of random tracking. A man threading his way through a crowd is looking ahead for the next opening as he passes between the first people.

This performance alone, and the delay coupled with it, make no particular demands on memory. Furthermore, the evidence suggests that any forced delay beyond the necessary processing time is likely to be injurious to the performance. It must also be clear that the task needs to be of such a kind that advanced information is available. An example of this kind is tracking a line with random movement. A moving point on a line is seen through a window, and a pointer is adjusted so that it will be on the line. If there is no coherence in the movement of the line, no advanced information, and a delay in the execution of the tracking, then errors will be random and equal in range to the amplitude of the movement. But if there is a window ahead of the point tracked equal to or slightly greater than the tracking delay, performance is optimum.

From the point of view of memory the more interesting cases are those in which the sequence is not random. Speech, and in particular meaningful speech, represents just such a sequence. For such material the eye-voice span, to take one instance, lengthens out a great deal. The effect is perfectly familiar to anyone who reads aloud. The phrasing, the emphasis, the melodic pattern is anticipated by the reader whose eye ranges well in advance of his voice. A listener can almost gauge the span of the reader by the periodicity of his speech, to some extent by the points at which he pauses. Furthermore, these periods are a rather direct function of the informational (*i.e.* surprise) value of the text.

I believe that fairly direct studies have been attempted on this span in typewriting. The technical difficulties are considerable and no published results are available. But the general results that have been reported are the same. For more meaningful material the span increases over what it is for material with no sequential coherence (*).

To sum up, the human organism has a short-term active memory limited to a matter of a few seconds duration, to perhaps ten « chunks » of information, and to moderate complexity, perhaps 10 bits per chunk. This active memory is of the greatest importance in exploiting the sequential properties of information that comes to us and it is of equal importance in organizing the sequence of actions that a person carries out.

In this connection, I should like to point out that there is a large area, as yet little explored, in which there are a related set of problems. These have to do with the highly complex motor skills of a man (or a race horse!). In speaking, in playing tennis, in dancing, in the skilled movements of a carpenter, a mason, or a musician, in all these cases the pattern of the movement is formed within the central nervous system under only rather general and supervisory control of the peripheral sense organs. LASHLEY has recently discussed this problem. All I can say here is that its analysis lies ahead of us and contributes very little just now to our problem.

5. – Learning.

The problems of learning and of long time memory are of such complexity that I shall not take them up here.

(*) There exists a related problem with regard to the spatial coherence of a visual pattern but almost nothing at all is known about this.

6. – Man as a channel.

This brings me to a point where I should like to venture a few rather general statements about the human organism as a processor of information. Throughout this discussion I have emphasized the fact that a man has a rather low capacity for handling simultaneous information. At no point do his achievements appear to be particularly great. Furthermore, it is my impression that physiological studies of the central nervous system suggest the same conclusion. Perhaps my colleagues here will dispute this. I can only report to you that when I first started making records of potentials in the human brain more than thirty years ago, we thought we might be missing something because our electrodes were so gross. And yet what we found was a considerable correlation among signals from electrodes spaced over several millimeters.

There are a number of alternatives open to us if we speculate about how men achieve so much with a mechanism of this kind. One possibility lies in the operation of what I have referred to as « set ». This is a mechanism that operates broadly to re-program the human computer, to rearrange the functions played by his various parts so that the greatest resolving power is brought to bear at the point where it is most needed. McKay made a somewhat similar point several years ago when he suggested that high precision was attained by making successive adjustments, each time increasing the sensitivity of the feed-back loop as the adjustment became closer. In McKay's scheme the hierarchical arrangement for changing the parameters of the system had to do only with precision. But the same argument applies, for example, in the comprehension of language, and perhaps in memory and recall. A slight change in the set that a man has can make all the difference between productive thinking and a stupid response. The suggestion is that by directing attention to a limited part of a problem, the informational requirements are reduced to manageable proportions.

A very different instance that argues in the same direction is the relation of eye movements to detailed vision. As a camera the eye is impossibly bad. I am sure that almost none of you are aware how extremely limited is the field in which you have sharp vision. If you fixate any point on this page and try to read the letters in a line two or three centimeters away, you will find that you cannot do it. And yet you stand on the shore of the lake here and believe that you take in all the details of the glorious scene before you, the fisy jumping, the sail on the horizon, the jewel-like roofs of the houses in the distance. Actually the « seeing » of a scene like this requires a hundred darting glances with the eyes focussed each time on one small part of the total picture. Lying back of your eyes there must be a switching system that keeps up with this changing flow of fine details and their varying relations, a mechanism that sorts the successive impressions into just their proper place and combines them with all that you know about lakes and mountains and how wonderful it is to be alive. Once more we see that the unusual feature of the organization of the nervous system is its ability to take a low capacity channel and to program it so that you achieve a high capacity output. There is, of course, some trading of space for time, and a requirement for an active memory. But even here the point to remember is that there is no fixed program, and the amount of recoding that is done depends upon what needs to be done.

One respect in which the natural system is different from an artificial system is the way in which it handles the problem of error. As von Neumann remarked some years ago, the test of a large computer is its ability to handle its own errors. This has led in the direction of building artificial machines of steadily increasing reliability, and of error-detecting devices that keep error at a very low level of probability. The human nervous system obviously does not work this way. Any one part of the system is relatively noisy, but the likely output would appear to be subject to constant check and correction. In this way, the initial information which has a low level of likelihood attached to it gradually undergoes refinement until the final result may be quite exact.

Another difference may lie in the employment of more kinds of elements than we now recognize. We tend to think of nervous activity as made up of a single mode of response. In contrast with this idea, I am reminded of the remark in von Neumann's (1957) posthumous book that it generally becomes inefficient to build the memory of a large computer out of active elements. Perhaps there are several kinds of control in the nervous system that more nearly resemble passive elements. There is talk about « state » variables in both psychology and physiology today, but it is too early to see just where this kind of thinking leads.

Any connection of this kind of speculation with information theory is pretty remote, as I should like to be the first person to admit. But then the test of information theory, as of any other theory, is in its usefulness in suggesting new facts, and new ways of looking at old facts. This certainly has happened in psychology. Many very ancient psychological problems have a new appearance as a result of the subject this Course has been discussing. The only trouble I have is in knowing what information theory is — is it physics, or mathematics, or game theory, or perhaps philosophy?

BIBLIOGRAPHY

J. G. Beebe-Center, M. S. Rogers and D. N. O'Connell: *Journ. Psychol.*, **39**, 157 (1955).

L. Carmichael and W. Dearborn: *Reading and visual fatigue* (Boston, 1947).

A. D. de Groot: *Het Dencken van den Schaker* (Amsterdam, 1946).

C. W. Eriksen and H. W. Hake *Journ. Exp. Psychol.*, **49**, 323 (1955).

W. R. Garner: *Journ. Exp. Psychol.*, **46**, 373 (1953).

H. W. Hake and W. R. Garner: *Journ. Exp. Psychol.*, **42**, 358 (1951).

J. R. M. Hayes: *Memory span for several vocabularies as a function of vocabulary size.* Quarterly Progress Report of Acoustic Laboratory, Massachusetts Institute of Technology (Jan-Jun. 1952).

W. E. Hick: *On the rate of gain in information*, in *Journ. Exp. Psychol.*, **44**, 11 (1952).

R. Hyman: *Journ. Exp. Psychol.*, **45**, 188 (1953).

E. T. Klemmer: *A further study of information transmission with matrix patterns.* Operational applications laboratory report. No. AFCRC-IN-57-I, Sept. 1957. (Available as ASTIA document AD110066).

E. T. Klemmer and F. C. Frick: *Journ. Exp. Psychol.*, **45**, 15 (1953).

E. T. Klemmer and P. F. Muller jr., *The rate of handling information. Key pressing responses to light patterns.* Human Factors Operations Research Laboratories Memo Report, N° 34, (March. 1953).

G. A. Miller: *Psychol. Rev.*, **63**, 81 (1956).

G. A. Miller: *Hu nan memory and the storage of information*, in *IRE Trans.*, Vol. IT-2, 129 (1956).

G. A. Miller and J. Selfridge: *Amer. Journ. Psychol.*, **63**, 176 (1950).

I. Pollack: *Journ. Acoust. Soc. Amer.*, **25**, 745 (1953).

I. Pollack and L. Ficks: *Journ. Acoust Soc. Amer.*, **26**, 155 (1954).

H. Quastler (ed.): *Information Theory in Psychology. Problems and Methods.* (Glencoe, 1956).

H. Quastler and B. Brabb: *Human performance in information transmission* V. *The Force of Habit.* Control Systems Laboratory Report N° R-70. (Jan. 1956).

J. von Neumann: *The Computer and the Brain* (New Haven, 1958).

SEMINARS

On the Coding Theorem and its Converse for Finite-Memory Channels (*).

A. Feinstein

Stanford University - Stanford, Cal.

1. – Introduction.

A central problem of information theory is the following:

Let X and Y be sets, consisting of finite numbers of elements denoted by x and y, respectively. For each $x \in X$ let $p(\ |x)$ denote a probability distribution on Y. For any positive integer n, let X^n and Y^n denote the product spaces $\prod_{i=1}^{n} \times X_i$ and $\prod_{i=1}^{n} \times Y_i$, where $X_i = X$ and $Y_i = Y$. We will denote the elements of X^n by u, and of Y^n by v. For each $u \in X^n$ we define a probability distribution $p(\ |u)$ on Y^n according to $p(\ |u) = p(\ |x_1) \times \ldots \times p(\ |x_n)$ where $u = (x_1, \ldots, x_n)$. For a fixed e, $0 \leqslant e < 1$, let $N(n, e)$ be the largest integer for which there exists a set $u_1, \ldots, u_{N(n,e)}$ of elements in X^n and disjoint sets $B_1, \ldots, B_{N(n,e)}$ in Y^n such that $p(B_i|u_i) \geqslant 1 - e$ for $i = 1, \ldots, N(n, e)$. What can be said about the behaviour of $N(n, e)$? The following results have been known for several years.

Coding theorem for discrete channel without memory. – There exists a constant $C \geqslant 0$ (which is in general non-zero) such that for any e, $0 < e < 1$, and H, $0 \leqslant H < C$ there is an $n(e, H)$ such that $N(n, e) \geqslant 2^{nH}$ for all $n \geqslant n(e, H)$.

Weak converse. – The statement of the coding theorem is not true for any $H > C$. Specifically we have $\log N(n, e) \leqslant (nC+1)/(1-e)$ for all n.

Quite recently Wolfowitz [4] has obtained a sharper estimate for $N(n, e)$ as follows:

(*) This work was carried out at Stanford University, Stanford, Cal. and was supported in part by the Office of Naval Research, United States Navy.

Strong converse. – For any e, $0 \leqslant e < 1$, we have $\limsup_n (1/n) \log N(n, e) \leqslant C$.

The constant C is called the channel capacity. All logarithms here and henceforth are taken to the base 2.

The coding theorem and the strong converse may be summed up by the assertion $\lim_{n\to\infty} (1/n) \log N(n, e) = C$, $0 < e < 1$. However, for the purpose of generalizing the problem under consideration, it is best to consider these three results separately.

The case $e = 0$ is singular, and appears to offer greater difficulties than the case $e > 0$. That is, while it is easily shown that $\lim_{n\to\infty} (1/n) \log N(n, 0) = C_0$ exists, a simple algorithm for determining C_0, even in some of the simplest non-trivial cases, is not known.

The triple X, Y, $p(\ |x)$ is said to define a discrete channel without memory, and C is called its capacity. The capacity is determined as follows: given a probability distribution $p(\ )$ on X, we can define a distribution on $X \times Y$ by $p(x, y) = p(x)p(y|x)$ and a distribution on Y by $p(y) = \sum_X p(x, y)$. We define $H(X) = -\sum_X p(x) \log p(x)$, $H(Y) = -\sum_Y p(y) \log p(y)$, and $H(X, Y) = -\sum_{X.Y} p(x, y) \log p(x, y)$. in which we take $0 \log 0 = 0$. Then the quantity $R_p = H(X) + H(Y) - H(X, Y)$ is non-negative, and $C = \max_p R_p$. R_p is called the rate of the channel with respect to the input distribution $p(\ )$ on X; the existence of $\max_p R_p$ follows from a simple continuity argument (*).

The situation can be generalized in various directions; that which will interest us here is the following:

Let X, Y be as before, and let $X^I = \prod_{-\infty}^{\infty} \times X_i$, $Y^I = \prod_{-\infty}^{\infty} \times Y_i$, where $X_i = X$ and $Y_i = Y$. For each element $x_\infty \in X^I$ let $\nu(\ |x_\infty)$ be a probability measure on the Borel field $\mathcal{F}_Y$ generated by the cylinder sets in Y^I which satisfies the following conditions.

1) For any cylinder set $S \subset Y^I$, $\nu(S|x_\infty)$ is measurable with respect to the Borel field $\mathcal{F}_X$ generated by the cylinder sets in X^I.

2) $\nu(\ |x_\infty)$ is stationary in the sense that for any cylinder $S \subset Y^I$ we have $\nu(TS|Tx_\infty) = \nu(S|x_\infty)$, where T is the shift transformation defined on X^I (and Y^I similarly) by $(Tx_\infty)_n = (x_\infty)_{n+1}$, where $(\ )_n$ denotes the n-th component of the term within the brackets.

3) $\nu(\ |x_\infty)$ is non-anticipating; *i.e.*, if $S \in \mathcal{F}_Y$ is such that there is a fixed t for which $(..., y_{-1}, y_0, y_1, ...) \in S$ implies that $(,.. y'_{-1}, y'_0, y'_1, ...) \in S$ if

(*) For a complete treatment of the various results which are stated here and further on, see FEINSTEIN [1].

$y'_i = y_i$ for all $i \leqslant t$, then $\nu(S|x_\infty) = \nu(S|x'_\infty)$ whenever $x'_i = x_i$ for all $i \leqslant t$. The triple X, Y, $\nu(\ |x_\infty)$ is said to define a discrete channel with memory.

Let $\mu(\ )$ be a stationary probability measure on $\mathcal{F}_X$. Then

$$\omega(A \times B) = \int_A \nu(B|x_\infty)\mu(\mathrm{d}x_\infty)\,,$$

for any cylinders $A \subset X^I$ and $B \subset Y^I$ defines a stationary probability measure $\omega(\ )$ on the space $X^I \times Y^I \sim (X \times Y)^I$, and $\eta(B) = \omega(X^I \times B)$ for any cylinder $B \subset Y^I$ defines a stationary probability measure $\eta(\ )$ on $\mathcal{F}_Y$. Now each element $(u, v) \in X^n \times Y^n$ defines a cylinder set in $(X \times Y)^I$, namely that consisting of all pairs (x_∞, y_∞) for which the 1-st, 2-nd, ..., n-th components of x_∞ and y_∞ respectively are $x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots, y_n$, where $u = (x_1, \dots, x_n)$ and $v = (y_1, \dots, y_n)$. Thus the measure $\omega(\ )$ on $(X \times Y)^I$ induces a probability distribution $\omega_n(\ )$ on $X^n \times Y^n$. Similarly the measures $\mu(\ )$ and $\eta(\ )$ on X^I and Y^I define distributions $\mu_n(\ )$ and $\eta_n(\ )$ on X^n and Y^n, and $\mu_n(u) = \omega_n(u \times Y^n)$, $\eta_n(v) = \omega_n(X^n \times v)$. If we define $R_{\mu n} = H(X^n) + H(Y^n) - H(X^n, Y^n)$, then $\lim\limits_{n\to\infty} (1/n)R_{\mu n} = R_\mu$ exists. The quantity $C_s = \operatorname*{lub}\limits_{\mu} R_\mu$, where lub is taken over all stationary probability measures $\mu(\ )$, is called the stationary capacity of the channel X, Y, $\nu(\ |x_\infty)$. The quantity $C_e = \operatorname*{lub}\limits_{\mu} R'_\mu$, where R'_μ denotes that the measure $\omega(\ )$ is ergodic, is called the ergodic capacity of the channel. Clearly $C_e \leqslant C_s$ if C_e exists; however, there are channels for which $\omega(\ )$ is not ergodic for any choice of $\mu(\ )$ (FEINSTEIN [1], p. 97).

For the general discrete channel with memory, as defined above, no reasonable analogue of the coding theorem or its converses is known. For channels with finite memory, however, an analogue of the coding theorem does hold.

A discrete channel with memory is said to have finite memory if there exists a non-negative integer m such that:

m.1. For any cylinder $[y_t, \dots, y_r]$ in Y^I we have $\nu([y_t, \dots, y_r]|x_\infty) = \nu([y_t, \dots, y_r]|x'_\infty)$ whenever $x'_i = x_i$, $i = t - m, \dots, r$.

m.2. For any two cylinders $[y_i, \dots, y_j]$ and $[y_k, \dots, y_n]$ such that $j + m < k$ we have $\nu([y_i, \dots, y_j] \cap [y_k, \dots, y_n]|x_\infty) = \nu([y_i, \dots, y_j]|x_\infty)\nu([y_k, \dots, y_n]|x_\infty)$ for all $x_\infty \in X^I$.

The smallest integer m for which both *m*.1 and *m*.2 hold is called the memory of the channel.

For a channel with finite memory m it is evident that $\nu([y_{m+1}, \dots, y_n]|\cdot|[x_1, \dots, x_n])$ is well defined for any $n > m$. Thus $\nu(\ |x_\infty)$ defines, for any $n > m$, a probability distribution on Y^{n-m} for every element $u = (x_1, \dots, x_n)$ in X^n, which we will denote by $\nu(\ |u)$. Specifically, we define $\nu(\ |u)$ by

$\nu(y_1, \ldots, y_{n-m}|u) = \nu([y'_{m+1}, \ldots, y'_n]|[x_1, \ldots, x_n])$ where $y'_{m+1} = y_1, \ldots, y'_n = y_{n-m}$. Then the following is known:

Coding theorem for discrete finite-memory channels. – For any e, $0 < e < 1$) and H, $0 \leqslant H < C_e$, there exists an $n(e, H)$ such that for any $n \geqslant n(e, H)$ there exist elements $u_1, \ldots, u_N$ in X^n and disjoint sets $B_1, \ldots, B_N$ in Y^{n-m} such that $\nu(B_i|u_i) \geqslant 1 - e$ and $N \geqslant 2^{nH}$.

Since it can be shown that for a finite-memory channel, the ergodicity of $\mu(\ )$ implies that of $\omega(\ )$, C_e exists and is in general non-zero.

2. – Channel capacity and the coding theorem.

In this Section we will give a new definition of capacity for a discrete finite-memory channel, and derive the coding theorem and its weak converse in terms of this capacity. We will also show that this capacity is not less than C_s; in the following Section we will see that actually $C_e = C_s = C$.

We have seen that for each $n > m$, $\nu(\ |u)$ is a well defined probability distribution on Y^{n-m} for every $u \in X^n$. Thus X^n, Y^{n-m}, $\nu(\ |u)$ define a discrete channel without memory; let C_n denote its capacity.

Definition. – By the capacity of the discrete finite-memory channel X, Y, $\nu(\ |x_\infty)$ we will mean the quantity $C = \operatorname{lub}_{n>m} C_n/n$.

Theorem 1. $C = \lim_{n\to\infty} C_n/n < \infty$.

Proof. We will show that $C_{i+j} \geqslant C_i + C_j$, $i, j > m$. Therefore $-C_i$ is a subadditive function, and so $\lim_{i\to\infty} (-C_i/i) = \operatorname{glb}_{i>m} (-C_i/i) = -\operatorname{lub}_{i>m} C_i/i$. To show that $C_{i+j} \geqslant C_i + C_j$, let $p_i(\ )$ and $p_j(\ )$ be probability distributions on X^i and X^j respectively for which the capacities C_i and C_j are achieved. Then $[p_i \times p_j](\ )$ defines a probability distribution on X^{i+j}; let R_{i+j} be the rate of X^{i+j}, Y^{i+j-m}, $\nu(\ |u)$ with respect to $p[_i \times p_j](\ )$. Let us apply now the data process on Y^{i+j-m} which identifies $(y_1, \ldots, y_{i+j-m})$ and $(y'_1, \ldots, y'_{i+j-m})$ if $y'_k = y_k$ for $k \neq i - m + 1, \ldots, i$ and let R'_{i+j} be the rate after data processing. Then $R_{i+j} \geqslant R'_{i+j}$; but by virtue of $m.2$ and the product form of the input distribution $[p_i \times p_j](\ )$ it follows easily that $R_{i+j} = C_i + C_j$. Since $C_{i+j} \geqslant R_{i+j}$, the proof is complete. As for $C < \infty$, we have $C_n \leqslant D^n$ where D^n is the number of elements in X^n; hence $C \leqslant \log D$.

Theorem 2. $C_s \leqslant C$.

Proof. Let $\mu(\ )$ be a stationary probability measure on $\mathcal{F}_X$, and let R_{n+m} be the rate of the memoryless channel X^{n+m}, Y^n, $\nu(\ |u)$ with respect to the distribution defined on X^{n+m} by $\mu(\ )$. Then $C_{n+m} \geqslant R_{n+m}$. On the other hand, if we contract (see Appendix I) with respect to the first m components of X^{n+m}, then the rate is precisely $R_{\mu n}$, and therefore $R_{\mu n} \leqslant R_{n+m} \leqslant C_{n+m}$ for all $n \geqslant 1$. Therefore $R_\mu = \lim\limits_{n\to\infty} R_{\mu n}/n \leqslant \lim\limits_{n\to\infty} C_{n+m}/n = \lim\limits_{n\to\infty} C_{n+m}/(n+m) = C$, and so $C_s = \operatorname*{lub}\limits_{\mu} R_\mu \leqslant C$.

Coding theorem for discrete finite-memory channels. – Let X, Y, $\nu(\ |x_\infty)$ be a discrete finite-memory channel with capacity $C > 0$. Then for any e, $0 < e < 1$ and H, $0 \leqslant H < C$ there is an $n(e, H)$ such that for any $n \geqslant n(e, H)$ we can find elements $u_1, \dots, u_N$ in X^n and disjoint sets $B_1, \dots, B_N$ in Y^{n-m} such that $\nu(B_i|u_i) \geqslant 1 - e$, $i = 1, \dots, N$ and $N \geqslant 2^{nH}$.

Proof. Since $C = \lim\limits_{n\to\infty} C_n/n$ there is a k for which $C_k > kH$. Choose an H' satisfying $kH < H' < C_k$; then by the coding theorem for the memoryless channel X^k, Y^{k-m}, $\nu(\ |u)$ there is an $n_k(e, H')$ such that for any $s \geqslant n_k(e, H')$ there is a set $w_1, \dots, w_N$ in X^{ks} and disjoint sets $B_1, \dots, B_N$ in $Y^{(k-m)s}$ such that $p(B_i|w_i) \geqslant 1 - e$, $i = 1, \dots, N$ and $N \geqslant 2^{sH'}$ where $p(\ |w) = \nu(\ |u_1) \times \dots \times \nu(\ |u_s)$ with $w = (u_1, \dots, u_s)$. Now to each element $y_1, \dots, y_{(k-m)s}$ in $Y^{(k-m)s}$ we associate a set $\varphi(y_1, \dots, y_{(k-m)s})$ in Y^{ks-m}, defined by $\varphi(y_1, \dots, y_{(k-m)s}) = (y_1, \dots, y_{k-m}) \times Y^m \times (y_{k-m+1}, \dots, y_{2(k-m)}) \times Y^m \times \dots \times (y_{(k-m)(s-1)+1}, \dots, y_{(k-m)s})$. If we define $B_i' = \varphi(B_i)$, $i = 1, \dots, N$, then the B_i' are disjoint sets in Y^{ks-m}, and by $m.2$ it follows (*) that $p(B_i|w_i) = \nu(B_i'|w_i) \geqslant 1 - e$, $i = 1, \dots, N$, and $N \geqslant 2^{sH'} > 2^{ksH}$, which proves the coding theorem for all n of the form ks for $s \geqslant n_k(e, H')$. Let us assume $n_k(e, H')$ taken so large that

$$\frac{n_k(e, H')}{n_k(e, H') + 1} \geqslant \frac{kH}{H'}.$$

For any $n \geqslant kn_k(e, H')$ set $n = s_nk + r$, where $0 \leqslant r < k$. For $n' = s_nk$ the theorem is proved; let $u_1, \dots, u_N$ and $B_1, \dots, B_N$, $N \geqslant 2^{s_nH'}$ be the corresponding elements in $X^{n'}$ and sets in $Y^{n'-m}$ respectively. Let z_0 be any fixed element in X^r; then (z_0, u_i) and $B_i' = Y^r \times B_i$, $i = 1, \dots, N$ defines elements w_i in X^n and disjoint sets B_i' in Y^{n-m}. By $m.1$ follows that $\nu(B_i'|w_i) = \nu(B_i|u_i)$ for all i. Since $N \geqslant 2^{s_nH'}$ and $s_nH'/n > \{(n_k(e, H'))/(n_k(e, H') + 1)\}(H'/k) \geqslant H$, we have $N > 2^{nH}$ which completes the proof.

Weak converse. – Let $N(n, e)$ be the largest integer for which we can find elements $u_1, \dots, u_{N(n,e)}$ in X^n and disjoint sets $B_1, \dots, B_{N(n,e)}$ in Y^{n-m} such that $\nu(B_i|u_i) \geqslant 1 - e$, $i = 1, \dots, N(n, e)$, where e satisfies $0 \leqslant e < 1$. Then $\log N(n, e) \leqslant (nC + 1)/(1 - e)$.

Proof. It is shown in FEINSTEIN [1] that the existence of elements $u_1, \dots, u_{N(n,e)}$ and sets $B_1, \dots, B_{N(n,e)}$ having the stated properties implies that $C_n \geqslant \log N(n, e) - e \log N(n, e) - 1$. But $C \geqslant C_n/n$, from which the desired results follows.

(*) See also FEINSTEIN [1], p. 104.

3. – Equality of C_s, C_e and C.

We have seen above that $C_s \leqslant C$. It follows from the definition of C^e that $C_e \leqslant C_s$. In order to show that $C_s = C_e = C$, we will show that $C_e \geqslant C$· This will be accomplished by constructing, for each $j > m$, an ergodic probability measure $\mu_j(\ )$ on X^I such that $R_{\mu_j} \geqslant C_j/j$. Now it is easily shown (cf. FEINSTEIN [1], p. 99-103) that for a finite-memory channel the ergodicity of $\mu(\ )$ implies that of $\omega(\ )$. Hence $C_e \geqslant \operatorname{lub}_j R_{\mu_j} \geqslant \operatorname{lub}_j C_j/j = C$.

The construction of $\mu_j(\ )$ is based on the following considerations (*). Let $p(\ )$ be, for some fixed s, a probability distribution on X^s. We define a probability measure $q(\ )$ on $\mathcal{F}_X$ as follows: for any integer $m > 0$ we put $q([x_{-ms+1}, \dots, x_{ms}](= p(x_{-ms+1}, \dots, x_{-(m-1)s}) \dots p(x_{(m-1)s+1}, \dots, x_{ms})$. Since any cylinder set in X^I is the union of finitely many disjoint cylinders of the type for which we have defined $q(\ )$, it follows readily that $q(\ )$ is well defined on $\mathcal{F}_X$. For arbitrary $p(\ )$, $q(\ )$ will not in general be stationary with respect to the shift transformation T on X^I. However $q(\ )$ is evidently stationary with respect to T^s. We define $\bar{q}(A) = 1/s(q(A) + q(TA) + \dots + q(T^{s-1}A))$ for all $A \in \mathcal{F}_X$; then $\bar{q}(TA) = 1/s(q(TA) + q(T^2A) + \dots + q(T^sA)) = \bar{q}(A)$ since $q(T^s)A = = q(A)$. Thus $\bar{q}(\ )$ is a stationary probability measure on $\mathcal{F}_X$.

Lemma 1. The measure $\bar{q}(\ )$ is ergodic with respect to T.

Proof. It is well known (see *e.g.* FEINSTEIN [1], p. 99-102) that for the ergodicity of a probability measure $\mu(\ )$ on $\mathcal{F}_X$ it is sufficient (and indeed necessary also) that $\lim\limits_{n\to\infty} 1/n \sum\limits_{i=0}^{n-1} \mu(T^{-i}A \cap B) = \mu(A)\mu(B)$ for all cylinders $A, B \subset X^I$. Due to the linearity of both sides with respect to A and B, it evidently suffices to verify this condition for A and B of the form $[x_{-ms+1}, \dots, x_{ms}]$. Now $\bar{q}(T^{-i}A \cap B) = 1/s\,[q(T^{-i}A \cap B) + q(T^{-i+1}A \cap TB) + \dots + q(T^{-i+s-1}A \cap T^{s-1}B)]$. From the definition of $q(\ )$ it easily follows that when $i \geqslant (2m+1)s - 1$ we have $q(T^{-i}A \cap B) = q(T^{-i}A)\,q(B)$, $q(T^{-i+1}A \cap TB) = q(T^{-i+1}A)\,q(TB)$, ... , $q(T^{-i+s-1}A \cap T^{s-1}B) = q(T^{-i+s-1}A)\,q(T^{s-1}B)$. Since by virtue of the stationary of $q(\ )$ with respect to T^s, $q(T^{-i}A)$, $q(T^{-i+1}A)$, ..., $q(T^{-i+s-1}A)$ all run cyclically through the values $q(A)$, $q(TA)$, ..., $q(T^{s-1}A)$ as i runs over all positive integers, while the value of $\lim\limits_{n\to\infty} 1/n \sum\limits_{i=0}^{n-1} q(T^{-i}A \cap B)$ is independent of the value of $q(T^{-i}A \cap B)$ for any finite number of values of i, it follows that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} q(T^{-i}A \cap B) = \frac{1}{s}[q(A) + q(TA) + \dots + q(T^{s-1}A)]q(B)\,,$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} q(T^{-i+l}A \cap TB) = \frac{1}{s}[q(A) + \dots + q(T^{s-1}A)]q(TB)\,,$$

(*) The costruction of $\mu_j(\ )$ was suggested by a result (Theorem 7) of J. NEDOMA [2].

etc. Hence

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\overline{q}(T^{-i}A\cap B)=\frac{1}{s^2}[q(A)+\dots+q(T^{s-1}A)][q(B)+\dots+q(T^{s-1}B)]=\overline{q}(A)\overline{q}(B)\,,$$

and so $\overline{q}(\;)$ is ergodic.

Lemma 2. Let $\beta(\;)$ be a probability measure on $\mathcal{F}_X$ which is stationary with respect to T^s. Then the quantity

$$H_\beta(X)=\lim_{n\to\infty}-\frac{1}{n}\sum_{x_{1+i},\dots,x_{n+i}}\beta([x_{1+1},\dots,x_{n+i}])\log\beta([x_{1+i},\dots,x_{n+i}])\,,$$

exists and is independent of i.

Proof. The demonstration is a simple adaptation of the usual one for stationary $\beta(\;)$. For each fixed r, the sequence $H(X_r|H_{r-1})$, $H(X_r|X_{r-1},X_{r-2})$, ... is known to be non-increasing, and therefore $H_r=\lim_{j\to\infty}H(X_r|X_{r-1},\dots,X_{r-j})$ exists. Clearly the sequence H_1, H_2, ... is periodic with period s. Now

$$-\log\beta([x_{1+i},\dots,x_{n+i}])=-\log\beta([x_{1+i}])-$$
$$-\log\frac{\beta([x_{1+i},x_{2+i}])}{\beta([x_{1+i}])}-\dots-\log\frac{\beta([x_{1+i},\dots,x_{n+i}])}{\beta([x_{1+i},\dots,x_{n+i-1}])}\,.$$

Multiplying the right side by $\beta([x_{1+i},\dots,x_{n+i}])$ and summing over $x_{1+i},\dots,x_{n+i}$, we obtain $H(X_{1+i})+H(X_{2+i}|X_{1+i})+\dots+H(X_{n+i}|X_{n+i-1},\dots,X_{1+i})$. Let us single out the terms $H(X_{1+i})$, $H(X_{1+i+s}|X_{i+s},\dots,X_{1+i})$, $H(X_{1+i+2s}|X_{i+2s},\dots,X_{1+i})$,... Evidently $H(X_{1+i+s}|X_{i+s},\dots,X_{1+i})=H(X_{1+i}|X_i,\dots,X_{1+i-s})$, $H(X_{1+i+2s}|X_{i+2s},\dots,X_{1+i})=H(X_{1+i}|X_i,\dots,X_{1+i-2s})$, etc. Therefore $\lim_{j\to\infty}H(X_{1+i+js}|X_{i+js},\dots,X_{1+i})==H_{1+i}$. Similarly $\lim_{j\to\infty}H(X_{2+i+js}|X_{1+i+js},\dots,X_{1+i})=H_{2+i}$, and so on. Using the simple fact that the Césaro averages of a convergent sequence converge to the limit of the sequence, it follows easily that

$$\lim_{n\to\infty}\frac{1}{n}H(X_{1+i})+H(X_{2+i}|X_{1+i})+\dots+H(X_{n+i}|X_{n+i-1},\dots,X_{1+i})=$$
$$=\frac{1}{s}(H_{1+i}+H_{2+i}+\dots+H_{s+i})\,.$$

Since the sequence H_1, H_2, ... has period s, the average on the right is independent of i, which completes the proof.

The following result was pointed out by L. Breiman.

Lemma 3. Let $q_1(\,), \ldots, q_s(\,)$ be a set of probability measures on $\mathcal{F}_X$ such that

$$H_i(X) = \lim_{n\to\infty} -\frac{1}{n} \sum_{x_1, \ldots, x_n} q_i([x_1, \ldots, x_n]) \log q_i([x_1, \ldots, x_n]) \,,$$

exists, $i=1, \ldots, s$. Given a set $a_1, \ldots, a_s$, $a_i \geqslant 0$, $\sum_{i=1}^{s} a_i = 1$, let $\bar{q}(\,) = \sum_{i-1}^{s} a_i q_i(\,)$. Then

$$\lim_{n\to\infty} -\frac{1}{n} \sum_{x_1, \ldots, x_n} \bar{q}([x_1, \ldots, x_n]) \log \bar{q}([x_1, \ldots, x_n]) = \sum_{i=1}^{s} a_i H_i(X) \,.$$

Proof. We have

$$\bar{q}([x_1, \ldots, x_n]) \log \bar{q}([x_1, \ldots, x_n]) =$$

$$= \sum_{i-1}^{s} a_i q_i([x_1, \ldots, x_n]) \log \bar{q}([x_1, \ldots, x_n]) = a_1 q_1([x_1, \ldots, x_n])$$

$$\left[\log a_1 q_1([x_1, \ldots, x_n]) + \log\left\{1 + \sum_{j\neq 1} \frac{a_j q_j([x_1, \ldots, x_n])}{a_1 q_1([x_1, \ldots, x_n])}\right\}\right] + \ldots +$$

$$a_s q_s([x_1, \ldots; x_n])\left[\log a_s q_s\,([x_1, \ldots, x_n]) + \log\left\{1 + \sum_{j\neq s} \frac{a_j q_j([x_1, \ldots, x_n])}{a_s q_s([x_1, \ldots, x_n])}\right\}\right].$$

Now

$$\lim_{n\to\infty} -\frac{1}{n} \sum_{x_1, \ldots, x_n} a_i q_i([x_1, \ldots, x_n]) \log a_i q_i([x_1, \ldots, x_n]) =$$

$$= a_i \lim_{n\to\infty} -\frac{1}{n} \sum_{x_1, \ldots, x_n} q_i([x_1, \ldots, x_n]) \log q_i([x_1, \ldots, x_n]) = a_i H_i(X) \,, \qquad i = 1, \ldots, s \,.$$

Furthermore, the inequality $\log(1+x) \leqslant x \log e$ implies that for each $i=1, \ldots, s$

$$0 \leqslant a_i q_i([x_1, \ldots, x_n]) \log\left\{1 + \sum_{j\neq i} \frac{a_j q_j([x_1, \ldots, x_n])}{a_i q_i([x_1, \ldots, x_n])}\right\} \leqslant \sum_{j\neq i} a_j q_j([x_1, \ldots, x_n]) \log e \,.$$

Thus

$$0 \leqslant \lim_{n\to\infty} \frac{1}{n} \sum_{x_1, \ldots, x_n}^{i} a_i q_i([x_1, \ldots, x_n]) \log\left\{1 + \sum_{j\neq i} \frac{a_j q_j([x_1, \ldots, x_n])}{a_i q_i([x_1, \ldots, x_n])}\right\} \leqslant$$

$$\leqslant \lim_{n\to\infty} \frac{1}{n} \sum_{x_1, \ldots, x_n}^{i} \sum_{j\neq i} a_j q_j([x_1, \ldots, x_n]) \log e \leqslant \lim_{n\to\infty} \frac{1}{n} \sum_{j\neq i} a_j \log e = 0 \,,$$

for each i, where $\sum_{x_1, \ldots, x_n}^{i}$ indicates summation over those $x_1, \ldots, x_n$ for which

$q_i([x_1, \ldots, x_n]) > 0$. Combining this with the preceding results establishes the lemma.

Theorem 3. $C_e = C_s = C$.

Proof. In view of Theorem 2 and the evident relation $C_e \leqslant C_s$, it suffices to show that $C_e \geqslant C$. Since $C = \lim\limits_{j\to\infty} C_j/j$ we can find, for any $\varepsilon > 0$, an r such that $C_r/r \geqslant C - \varepsilon$. Let $p(\;)$ be the probability distribution on X^r for which C_r is attained, and let $q(\;)$ and $\bar{q}(\;)$ be the probability measures on $\mathcal{F}_X$ derived from $p(\;)$ as defined preceding Lemma 1. Since $q(\;)$ is stationary with respect to T^r, it follows that the probability measures $\omega(\;)$ and $\eta(\;)$ defined by $\omega(A\times B) = \int\limits_A \nu(B\,|\,x_\infty)q(\mathrm{d}x_\infty)$ and $\eta(B) = \omega(X^I \times B)$ for any cylinders $A \subset X^I$, $B \subset Y^I$ are likewise stationary with respect to T^r on $(X\times Y)^I$ and Y^I respectively. Let $q_i(S) = q(T^i S)$ for $S \in \mathcal{F}_X$, $i = 0, \ldots, r-1$. Applying Lemma 2 to $q_i(\;)$, $\omega_i(\;)$, and $\eta_i(\;)$, it follows that the quantities $R_{q_i} = \lim\limits_{n\to\infty} R_{q_i n}/n$ exist and are independent of i; let their common value be R_q. By Lemma 3 as applied to $\bar{q}(\;)$, $\bar{\omega}(\;)$, and $\bar{\eta}(\;)$, it follows that $R_{\bar{q}} = R_q$, and from Lemma 1 and the definition of C_e follows then $C_e \geqslant R_{\bar{q}} = R_q$. We will now show that $R_q \geqslant C_r/r$. Now $R_q = \lim\limits_{d\to\infty} R_{qdr}/\mathrm{d}r$; but R_{qdr} is the rate of a channel (in the sense of Appendix I) whose input space is

$$\underbrace{X^r \times \ldots \times X^r}_{d \text{ factors}} \sim X^{dr}$$

and whose output space is

$$\underbrace{Y^r \times \ldots \times Y^r}_{d \text{ factors}} \sim Y^{dr}$$

Now apply the data process on Y^{dr} which, in each factor Y^r, identifies two elements $y_1, \ldots, y_r$ and $y'_1, \ldots, y'_r$ if $y_{m+1} = y'_{m+1}, \ldots, y_r = y'_r$. Let R'_{qdr} be the rate after data processing; then $R'_{qdr} \leqslant R_{qdr}$. But it follows from $m.2$ [1], and the product nature of $q(\;)$ that $R'_{qdr} = dC_r$. Hence $R_{qdr}/dr \geqslant C - \varepsilon$ for all d, and therefore $C_e \geqslant R_{\bar{q}} = R_q = \lim\limits_{d\to\infty} R_{qdr}/dr \geqslant C - \varepsilon$; but ε was arbitrary, hence $C_e \geqslant C$.

4. – Extension of a result of Wolfowitz.

In this Section we will consider a particular type of discrete finite-memory channel, for which we can establish the strong converse of the coding theorem. This channel, which has been studied recently by WOLFOWITZ [4], is defined as follows. Let $m > 0$ be a fixed integer, and for every $(x_1, \dots, x_{m+1}) \in X^{m+1}$ let $p(\ |x_1, \dots, x_{m+1})$ be a probability distribution on Y. Let $x_\infty = (\dots, x_{-1}, x_0, x_1, \dots)$; then we set

$$\nu(\ |x_\infty) = \dots \times p(\ |x_{-m}, \dots, x_0) \times p(\ |x_{-m+1}, \dots, x_1) \times \dots ,$$

where the «alignment» of the product measure is fixed by $\nu([y_0]|x_\infty) = p(y_0|x_{-m}, \dots, x_0)$. It is readily verified that $\nu(\ |x_\infty)$ defines a channel with memory m, for which $m.2$ is satisfied for $m = 0$.

We will prove the strong converse of the coding theorem for this channel, in a form which is clearly analogous to the memoryless case.

Strong converse. Let the channel X, Y, $\nu(\ |x_\infty)$ be as above, and let $C > 0$ be its capacity. Let e, H be fixed, such that $0 \leqslant e < 1$, $H > C$. Then it is not possible to find arbitrarily large n such that there exist elements $u_1, \dots, u_N \in X^n$ and disjoint sets $B_1, \dots, B_N$ in Y^{n-m} such that $\nu(B_i|u_i) \geqslant 1 - e$, $i = 1, \dots, N$, and $N \geqslant 2^{nH}$.

Proof. We proceed by contradiction. Given $H > C$ we choose integers $r > m$ and d so large that $((d-1)/d)((r-m)/r)H > C$; r and d are henceforth fixed. For $n > (d-1)(r-m)+m$ define k by $(k-1)(r-m)+m < n \leqslant k(r-m)+m$. Suppose that $n > (d-1)(r-m)+m$ is such that there exist elements $u_1, \dots, u_N \in X^n$ and disjoint sets $B_1, \dots, B_N$ in Y^{n-m} for which $\nu(B_i|u_i) \geqslant 1 - e$, $i = 1, \dots, N$, and $N \geqslant 2^{nH}$. Let $n' = k(r-m)+m \geqslant n$; then, as we have seen in the proof of the coding theorem, there exist sequences $u'_1, \dots, u'_N$ in $X^{n'}$ and disjoint sets $B'_1, \dots, B'_N$ in $Y^{n'-m}$ such that $\nu(B'_i|u'_i) \geqslant 1 - e$, $i = 1, \dots, N$. We now define a mapping φ which takes elements of $X^{n'}$ into elements of X^{kr} as follows:

$$\varphi(x_1, \dots, x_{n'}) =$$
$$= (x_1, \dots, x_r, x_{r-m+1}, x_{r-m+2}, \dots, x_{2r-m}, x_{2r-2m+1}, \dots, x_{n'-r+m}, x_{n'-r+1}, \dots, x_{n'}) .$$

In words, we begin by setting down the first r elements of $(x_1, \dots, x_{n'})$; then we repeat the last m, then set down the next $r - m$, then repeat the last m, then set down the next $r - m$, and so on, the process ending as soon as $x_{n'}$ has been reached. It is clear that distinct elements in $X^{n'}$ go into distinct elements in X^{kr} under the mapping φ. Now for each $w \in X^{kr}$, let $p(\ |w)$ be the probability distribution on $Y^{k(r-m)}$ defined by $p(|w) = \nu(\ |x_1, \dots, x_r) \times \dots \times \nu(\ |x_{(k-1)r+1}, \dots, x_{kr})$ where $\nu(\ |x_1, \dots, x_r)$ is, as usual, a probability distribution on Y^{r-m} for each $(x_1, \dots, x_r) \in X^r$. Now from the particular form of $\nu(\ |x_\infty)$ it follows that $\nu(\ |x_1, \dots, x_{n'})$ and $p(\ |\varphi(x_1, \dots, x_{n'}))$ are identical probability distributions on $Y^{n'-m} = Y^{k(r-m)}$. Let $w_i = \varphi(u_i)$, $i = 1, \dots, N$; then we have

$p(B'_i|w_i) \geqslant 1-e$, $i=1, \ldots, N$, $N \geqslant 2^{nH}$. But

$$H' = \frac{nH}{kr} = H\,\frac{n}{k(r-m)+m}\,\frac{k(r-m)+m}{kr} \geqslant H\,\frac{\mathrm{d}-1}{\mathrm{d}}\,\frac{r-m}{r} > C \geqslant \frac{C_r}{r},$$

so $N \geqslant 2^{nH} = 2^{krH'}$ where $rH' > C_r$. Now as n becomes arbitrarily large, k does also, and we therefore have a contradiction of the strong converse of the coding theorem for the memoryless channel defined by X^r, Y^{r-m}, $\nu(\ |x_1, \ldots, x_r)$, which completes the proof.

We may remark, in passing, that for the channels considered in this Section it is simple to obtain a necessary and sufficient condition for the vanishing of the capacity C. Indeed, since $C = \operatorname{lub}_{j>m} C_j/j$, it follows that $C=0$ implies $C_{m+1}=0$, which in turn implies (cf. FEINSTEIN [1], p. 32) that the set $\{p(\ |x_1, \ldots, x_{m+1})\}$, $(x_1, \ldots, x_{m+1}) \in X^{m+1}$ of probability distributions on Y consists of only one distinct number. Conversely, this last condition evidently implies that the set $\{\nu(\ |u)\}$, $u \in X^n$ of probability distributions on Y^{n-m} also consists of only one distinct member, which implies $C_n = 0$, for all $n > m$, and so $C=0$. Hence for the vanishing of C it is both necessary and sufficient that the set $\{p(\ |x_1, \ldots, x_{m+1})\}$, $(x_1, \ldots, x_{m+1}) \in X^{m+1}$ consist of only one distinct member. As a consequence of this result, the construction of examples of (both discrete and semi-continuous) finite-memory channels with non-zero capacity becomes trivial (*).

5. – Semi-continuous finite-memory channels.

In this Section we will discuss briefly the extent to which our results carry over from discrete finite-memory channels to semi-continuous finite-memory channels.

In essence, a semi-continuous channel is obtained from a discrete channel by replacing the finite space Y of the latter by an arbitrary space Z in which is defined a Borel field. Specifically, a semi-continuous channel without memory is defined by the usual space X, an arbitrary space Z in which is defined a Borel field $\mathcal{F}$, and, for each $x \in X$, a probability measure $p(\ |x)$ on $\mathcal{F}$. The rate of this channel with respect to a probability distribution $p(\ )$ on X may be defined by noticing that although $H(X, Y)$ and $H(Y)$ have no direct analogues in the semi-continuous case, their difference

$$H(X|Y) = H(X, Y) - X(Y) = -\sum_X \sum_Y p(x, y) \log p(x|y)$$

(*) This result replaces the discussion on p. 98 of FEINSTEIN [1], in which the measure $\eta(\ )$ is incorrectly (in general) asserted to be defined by a Markoff chain.

is well defined even in this more general situation. Indeed, $p(x|z)$ is well defined as a Radon-Nikodym derivative, and

$$H(X|Z) = -\sum_{X}\int_{Z} \log p(x|z)p(x, dz)\,,$$

is well defined; we put $R_p = H(X) - H(X|Z)$. The capacity is defined as before by $C = \max\limits_{p} R_p$, where again the existence of $\max\limits_{p} R_p$ follows from a continuity argument.

For a semi-continuous channel without memory, both the coding theorem and its weak converse are known to hold; the strong converse is at present undecided.

The definition of an arbitrary semi-continuous channel with memory proceeds in similar fashion; we replace Y by Z. However, a technical point arises in the definition of $\nu(\ |x_\infty)$. Let $Z^I = \prod\limits_{-\infty}^{\infty} \times Z_i$, $Z_i = Z$, and let $\mathcal{F}^I$ be the Borel field in Z^I which is determined by $\mathcal{F}$ in the usual manner. Let $\mathcal{F}^n$ be the Borel field of sets $S \in \mathcal{F}^I$ such that $(\ldots, z_{-1}, z_0, z_1, \ldots) \in S$ and $z_i' = z_i$, $i = -n, \ldots, n$ implies that $(\ldots, z_{-1}', z_0', z_1', \ldots) \in S$. Let $\mu(\ )$ be a set function defined on $\bigcup\limits_{n=1}^{\infty} \mathcal{F}^n$ which is a probability measure on each $\mathcal{F}^n$. Then it is not unrestrictedly true that $\mu(\ )$ can be extended to $\mathcal{F}^I$ as a probability measure.

Now in our results for the discrete case the only property of $\nu(\ |x_\infty)$ that was actually used was that it was defined on $\bigcup\limits_{n=1}^{\infty} \mathcal{F}^n$, and that it was a probability distribution on $\mathcal{F}^n$ for every n (and similarly for $\omega(\ )$ and $\eta(\ )$((*). The same is true here; it is sufficient to require only that $\nu(\ |x_\infty)$ is defined on $\bigcup\limits_{n=1}^{\infty} \mathcal{F}^n$ and is a probability measure of $\mathcal{F}^n$ for each n. It follows that for a given probability measure $q(\ )$ on $\mathcal{F}_X$, $\omega(\ )$ and $\eta(\ )$ are not necessarily measures on $\mathcal{F}_X \times \mathcal{F}^I$ or $\mathcal{F}^I$ respectively, but only on $\mathcal{F}_X \times \mathcal{F}^n$ and $\mathcal{F}^n$ for every n.

With these definitions it is easily seen that $C = \operatorname{lub}\limits_{j>m} C_j/j = \lim\limits_{j\to\infty} C_j/j$, and that the proofs of the coding theorem and its weak converse remain unaltered. The result $C_e = C_s = C$ requires certain modifications however. Here, the difficulty arises in the definition $R_\mu = \lim\limits_{n\to\infty} R_{\mu n}/n$ for the rate with respect to a given stationary probability measure $\mu(\ )$ on X^I; it is not known whether or not the limit R_μ actually exists. We can, however, avoid the question by

(*) We may remark that in the discrete case $\nu(|x_\infty)$ *can* be uniquely extended to $\mathcal{F}^I$ according to the Kolmogorov extension theorem; however, this fact was never required.

defining $R_\mu = \lim_n \sup R_{\mu n}/n$; then C_s can be defined as in the discrete case, and the proof of $C_s \leqslant C$ remains valid. In order to define C_e in a reasonable way, we have to impose some condition similar to ergodicity on $\omega(\ )$. A reasonable analogue to the discrete case would be that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \omega(T^{-i}A \cap B) = \omega(A)\omega(B)\,,$$

for any sets A, B in $\bigcup\limits_{n=1}^{\infty} \mathcal{F}_X \times \mathcal{F}^n$. Then it can be shown, just as in the discrete case, that for a finite-memory channel this condition is equivalent to the ergodicity of $\mu(\ )$. Clearly $C_e \leqslant C_s$ just as in the discrete case, and we show that $C_e = C_s = C$ by proving that $C_e \geqslant C$. Before doing so, we should point out that this proof is, at the moment, of no greater interest than an analytical exercise, for we do not know whether C_e, as defined here, enters into a statement of the coding theorem as it does in the discrete case.

The proof follows that in the discrete case with only minor variations. The first concerns the analogue of Lemma 3, or rather an analogue of the linearity of R as a function of $\mu(\ )$, which followed directly from Lemma 3 in the discrete case. Actually, the linearity of R is not essential; a slightly weaker result is sufficient. Indeed, from the proof of Lemma 3 the following is easily shown:

For each integer $r > 1$ these is a constant $F_r > 0$ with the following property: Let X, Y, $p(\ |x)$ be any discrete memoryless channel, let $p_1(\ ), \ldots, p_r(\ )$ be probability distributions on X, and let $R_1, \ldots, R_r$ be the corresponding rates of the channel. Set $\bar{p}(\ ) = 1/r[p_1(\ ) + \ldots + p_r(\ )]$; then $|R_{\bar p} - (1/r)\sum\limits_{i=1}^{r} R_i| \leqslant F_r$. An explicit value for F_r is readily obtained, but will not be needed.

The same result holds for semi-continuous channels without memory. Let $r = 2$ for simplicity; we set $p_i(x,\ \ ) = p_i(x)\,p(\ |x)$, $\eta_i(\ ) = \sum\limits_X p_i(x,\ \ )$, $i = 1, 2$; $\bar p(x) = \frac{1}{2}[p_1(x) + p_2(x)]$, $\bar p(x,\ \ ) = \bar p(x) p(\ \ |x)$, and $\bar\eta(\ ) = \frac{1}{2}[\eta_1(\ ) + \eta_2(\ )]$. Then

$$\bar H(X|Z) = -\sum_X \int_Z \log \bar p(x|z) \bar p(x, \mathrm{d}z) =$$

$$= -\frac{1}{2} \sum_X \int_Z \log \bar p(x|z) p_1(x, \mathrm{d}z) - \frac{1}{2} \sum_X \int_Z \log \bar p(x|z) p_2(x, \mathrm{d}z)\,.$$

But

$$\bar p(x|z) = \frac{\bar p(x, \mathrm{d}z)}{\bar\eta(\mathrm{d}z)} = \frac{p_1(x, \mathrm{d}z) + p_2(x, \mathrm{d}z)}{\eta_1(\mathrm{d}z) + \eta_2(\mathrm{d}z)} = \frac{p_1(x, \mathrm{d}z)}{\eta_1(\mathrm{d}z)} + \frac{p_2(x, \mathrm{d}z)}{\eta_1(\mathrm{d}z)} \Big/ 1 + \frac{\eta_2(\mathrm{d}z)}{\eta_1(\mathrm{d}z)}\,,$$

a.e. $\eta_1(\ )$, where the derivatives with respect to $\eta_1(\ )$ are, by the Lebesgue de-composition, well defined a.e. $\eta_1(\ )$. Further,

$$\frac{p_2(x,\,\mathrm{d}z)}{\eta_1(\mathrm{d}z)} = \frac{p_2(x,\,\mathrm{d}z)}{p_1(x,\,\mathrm{d}z)} \Big/ \frac{\eta_1(\mathrm{d}z)}{p_1(x,\,\mathrm{d}z)} = \frac{p_2(x,\,\mathrm{d}z)}{p_1(x,\,\mathrm{d}z)}\, \mathrm{p}_1(x\,|\,z)\ ,$$

a.e. $p_1(x,\ )$. Since a.e. $\eta_1(\ )$ implies a.e. $p_1(x,\ )$, we have

$$\overline{p}(x\,|\,z) = p_1(x\,|\,z)\left[1 + \frac{p_2(x,\,\mathrm{d}z)}{p_1(x,\,\mathrm{d}z)} \Big/ 1 + \frac{\eta_2(\mathrm{d}z)}{\eta_1(\mathrm{d}z)}\right],$$

a.e. $p_1(x,\ )$. We use this expression in

$$\frac{1}{2}\sum_X \int_Z \log \overline{p}(x\,|\,z)\, p_1(x,\,\mathrm{d}z)\ ,$$

and the same expression, but with the indices 1, 2 interchanged, in

$$\frac{1}{2}\sum_X \int_Z \log \overline{p}(x\,|\,z)\, p_2(x,\,\mathrm{d}z)\ .$$

From this point the proof follows that of the discrete case.

The demonstration of $C_e \geqslant C$ continues now as in the discrete case. Given $\varepsilon > 0$ we find r such that $C_r/r \geqslant C - \varepsilon$; we define $\overline{p}(\ ) = 1/r \sum_{i=0}^{r-1} q_i(\ )$, where $q_i(A) = q(T^i A)$, $i = 0, \ldots, r-1$, $A \in \mathcal{F}_X$. Now we have that $|R_{\overline{p}} - 1/r \sum_{i=0}^{r-1} R_{in}| \leqslant F_r$.

If we take $n = \mathrm{d}r$, then it follows, just as in the discrete case, that $R_{0n} \geqslant \mathrm{d}C_r$. To estimate R_{in}, $i = 1, \ldots, r-1$, we note that the channel (in the sense of Appendix I) which defines R_{in} is equivalent (in the sense of the isomorphism defined by T^i) to that defined by $\omega_0(\ )$, the family of cylinder sets of the form $[x_{1+i}, \ldots, x_{n+i}]$, and the space of all cylinder of the form $[z_{1+i}, \ldots, z_{n+i}]$. Now if we contract this channel with respect to $x_{1+i}, \ldots, x_r$ and $x_{dr+1}, \ldots, x_{dr+i}$, and apply the data process which identifies $z_{1+i}, \ldots, z_{n+i}$ and $z'_{1+i}, \ldots, z'_{n+1}$ if $z_{r+1} = z'_{r+1}, \ldots, z_{dr} = z'_{dr}$, we see that the resulting channel is equivalent, by virtue of the stationarity of $\omega_0(\ )$ with respect to T^r, to the channel which defines $R_{0(d-1)r}$. Thus $R_{in} \geqslant (d-1)C_r$, $i = 1, \ldots, r-1$ where $n = dr$, and we have

$$C_e \geqslant R_{\overline{q}} = \limsup_n \frac{R_{\overline{q}n}}{n} \geqslant \limsup_d \frac{R_{\overline{q}dr}}{\mathrm{d}r} \geqslant \limsup_d \frac{1}{\mathrm{d}r}[(\mathrm{d}-1)C_r - F_r] = \frac{C_r}{r} \geqslant C - \varepsilon\ ,$$

which completes the proof.

6. – Remarks.

Shortly after this work was completed, a paper appeared by I. P. TSAREGRADSKY [3], in which the quantity C is introduced and the relation $C_e = C_s = C$ is proved by methods similar to those used here.

At this Summer School we were also informed by F. L. STUMPERS that certain of our results have been obtained by J. WOLFOWITZ in a paper shortly to be published.

APPENDIX I

We have defined a discrete memoryless channel as consisting of the triple $X, Y, p(\ |x)$. Now given a probability distribution $p(x, y)$ on $X \times Y$, it is clear that we may put $p(x, y) = p(x)\, p(y|x)$, where $p(y|x)$ is, for each x, a probability distribution on Y, and is unique for each x for which $p(x) = p(x, Y) > 0$. The nonuniqueness is of no concern since the rate $R = H(X) + H(Y) - H(X, Y)$ is uniquely determined by $p(x, y)$; this point of view simplifies the discussion of the notion of a contraction of a channel.

Let $p(x, y)$ be a probability distribution on $X \times Y$, and let $A_1, \ldots, A_n$ be disjoint sets in X whose union is X. Then $p(A_i, y)$ is well defined, and $\mathscr{A}, Y, p(A_i, y)$, where $\mathscr{A} = \{A_1, \ldots, A_n\}$ defines a channel (or a family of channels) with a unique input distribution $p(A_i) = p(A_i, Y)$ on $\mathscr{A}$, having rate $R_c = H(\mathscr{A}) + H(Y) - H(\mathscr{A}, Y)$. The rate R_c we call the rate of the channel defined by $X, Y, p(x, y)$ after the contraction defined by the family $\mathscr{A}$, and $\mathscr{A}, Y, p(A_i, y)$ we call the contracted channel. To show that the process of contraction can never increase the rate of a channel, it suffices to observe that $X, Y, p(x, y)$ may equally well be considered as defining a channel with input space Y and output space X, since the rate $R = H(X) + H(Y) - H(X, Y)$ is symmetric in X and Y. From this point of view the contraction becomes a data process on the output of the «reversed» channel, in which form the non-increase of the rate is well known.

The notation commonly used in denoting cylinder sets in a product space is particularly convenient in this connection; if the input space of a channel consists of the family of cylinders $[x_1, \ldots, x_n]$, then by the contraction of this channel with respect to the component x_1 we mean the family of sets $[x_2, \ldots, x_n]$, each considered as a set of cylinders $[x_1, \ldots, x_n]$ by virtue of $[x_2, \ldots, x_n] = \bigcup\limits_{x_1} [x_1, \ldots, x_n]$.

The same considerations hold fer semi-continuous channels without memory, except for the proof that a contraction never increases the rate. This result can be established as follows: let R be the rate of the given channel and R_c its rate after a given contraction. For any $\varepsilon > 0$ there is a data process which reduces the contracted channel to a discrete one, and yet reduces its rate to a value R_{cd} such that $R_{cd} \geqslant R_c - \varepsilon$. Now if R_d is the rate of the original channel after this data process, then $R \geqslant R_d$. But $R_d \geqslant R_{cd}$, since we are now dealing with discrete channels. Thus $R \geqslant R_c - \varepsilon$ for any $\varepsilon > 0$, or $R \geqslant R_c$.

Appendix II

Let m be a non-negative integer, and let $\{a_i\}$, $i = m + 1, \ldots$ be an infinite sequence of finite terms such that $a_{i+j} \leqslant a_i + a_j$ for all $i, j > m$. Then $\{a_i\}$ is called a subadditive sequence, and we have that $\lim\limits_{i \to \infty} a_i/i$ exists and equals $\operatorname*{glb}\limits_{j > m} a_i/i = A$.

For the proof, assume first that $A > -\infty$; then for any $\varepsilon > 0$ there is an integer $s > m$ such that $a_s/s \leqslant A + \varepsilon$. For any $n > 2s$ we define an integer $k > 0$ according to $n = ks + r$ where $s \leqslant r < 2s$. Then $a_n \leqslant a_{ks} + a_r \leqslant ka_s + a_r$, and so $a_n/n \leqslant (ks/n)(a_s/s) + a_r/n$. Now as $n \to \infty$, $ks/n \to 1$, and we have $\limsup\limits_{n} a_n/n \leqslant a_s/s \leqslant A + \varepsilon$. Since ε is arbitrary, we have $\limsup\limits_{n} a_n/n \leqslant A$. But $a_n/n \geqslant A$, which imples that $\lim\limits_{n \to \infty} a_n/n = A$. The case $A = -\infty$ follows in similar fashion.

REFERENCES

[1] A. Feinstein: *Foundations of Information Theory*, (New York, 1958).

[2] J. Nedoma: *The Capacity of a Discrete Channel*, *Trans.* First Prague Conference on Information Theory, Statistical Decision Functions, Random Process (1956)·

[3] I. P. Tsaregradsky: *On the Capacity of a Stationary Channel with Finite Memory* (in Russian), in *Theory of Probability and its Applications*, **3**, 84 (1948).

[4] J. Wolfowitz: *The Coding of Messages Subject to Chance Errors*, in *Illinois Jour. of Math.*, **1**, 591 (Dec. 1957); *An Upper Bound on the Rate of Transmission of Messages*, in *Illinois Journ. of Math.*, **2**, 137 (March. 1958).

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3º Trimestre

Correlation Indices.

S. Watanabe

International Business Machines Co. Research Laboratory - Ossining, N.Y.

Suppose we are given n stochastic variables $y_1, y_2, \ldots, y_n$, each of which can take g discrete values or states, $v_1, v_2, \ldots, v_g$. These stochastic variables are not necessarily independent of one another, *i.e.*, they are somehow correlated. We are interested in introducing useful measures of such correlatioe in such a way that the following conditions are satisfied. 1) The measure is entirely independent of the values v_i $(i = 1, \ldots, g)$ assigned to the states, *i.e.*, it can be defined even if v_i are replaced by non-numerical symbols. 2) Thmeasure is always non-negative. 3) The total correlation $C^{(n)}$ $(\geqslant 0)$ is decomposed as $C^{(n)} = \sum_{r=1}^{n} T^{(r)}$, where $T^{(r)}$ is non-negative and measures, in a certain sense, the strength of the correlation peculiar to a subset of r variables taken out of the n variables. Claim is not made that the definitions proposed in this note are unique or the best, but it is intended that they will be meaningful and useful according to the usage.

We want to limit our discussion to two cases: 1) simultaneous, symmetric set of stochastic variables, 2) temporal, stationary sequence of stochastic variables. We shall first define the first case.

We are given n stochastic variables $y_1, y_2, \ldots, y_n$, each of which can take g discrete values, 1, 2, ..., g. The probability that the variables, $y_1, y_2, \ldots, y_n$, *simultaneously* take values $x_1, x_2, \ldots, x_n$, respectively, will be denoted by

$$p^{(n)}(x_1 = x_1,\ y_2 = x_2,\ \ldots,\ y_n = x_n) \geqslant 0\,, \tag{1}$$

with

$$\Big(\sum_{x=1}^{g}\Big)^n p^{(n)}(y_1 = x_1,\ y_2 = x_2,\ \ldots,\ y_n = x_n) = 1\,. \tag{1'}$$

We are speaking of probabilities here, considering an ensemble of similar sets of n variables. One can of course consider a time-ensemble of the same set, instead of the simultaneous ensemble of similar sets, but in this case one has to assume that there is no temporal correlation. By the *symmetric* set, we mean that the probability $p^{(n)}$ of (1) is invariant for an interchange of values of any two variables,

$$p^{(n)}(y_1=x_1,\, \dots,\, y_i=x_i,\, \dots,\, y_j=x_j,\, \dots,\, y_n=x_n) = \\ = p^{(n)}(y_1=x_1,\, \dots,\, y_i=x_j,\, \dots,\, y_j=x_i,\, \dots,\, y_n=x_n)\,. \tag{2}$$

The case where $p^{(n)}$ is not symmetric is discussed in a forthcoming article in the *IBM Journal of Research and Development.*

By summing up over possible values of any $(n-r)$ variables out of the n variables, one obtains the probability distribution of r variables. For instance,

$$p^{(r)}(y_1=x_1,\, \dots,\, y_r=x_r) = \sum_{x_r+1}^{g} \dots \sum_{x_n}^{g} p^{(n)}(y_1=x_1,\, \dots,\, y_n=x_n)\,. \tag{3}$$

Because of the symmetry, the value of $p^{(r)}$ is determined only by the *values* $x_1, \dots, x_r$ and is independent of the r *variables* which take these values. Thus we write $p^{(n)}$ of (1) and $p^{(r)}$ of (3), as functions of *values* only: $p^{(n)}(x_1, \dots, x_n)$ and $p^{(r)}(x_1, \dots, x_r)$. The p's are of course invariant for any permutation of their arguments.

The « information » function of r variables $(r=0, 1, \dots, n)$ is defined by

$$S^{(r)} = -\sum_{x_1}^{g} \dots \sum_{x_r}^{g} p^{(r)}(x_1, \dots, x_r) \log p^{(r)}(x_1, \dots, x_r)\,, \tag{4}$$

where $S^{(r)}$ is a function only of r. Of course, $S^{(0)}=0$.

We shall next define the case of temporal, stationary sequences. We are given an infinite, one-dimensional (*temporal*) series of stochastic variables: $(\dots, y_{-3}, y_{-2}, y_{-1}, y_0, y_1, y_2, y_3, \dots)$, such that any arbitrary segment of r *consecutive* variables has a unique and definite probability of having a given ordered set of values, say $(x_1, x_2, \dots, x_r)$. This means that

$$p^{(r)}(y_1=x_1,\, \dots,\, y_r=x_r) = p^{(r)}(y_{1+k}=x_1,\, \dots,\, y_{r+k}=x_r)\,, \tag{5}$$

where k is an arbitrary integer, positive or negative. For this reason, we shall write the probability $p^{(r)}$ of (5), simply $p^{(r)}$ $(x_1, \dots, x_r)$. What matters here is only that $x_1, \dots, x_r$ are the values of any r *consecutive* variables. Permutations of arguments are not allowed.

Formula (3) is valid also in this case with a specific proviso that $y_1, \dots, y_r, \dots, y_n$ are *consecutive* variables. The summation can also be made with respect to the first $(n-r)$ variables instead of the last $(n-r)$ variables. The r-variable information function $S^{(r)}$ can also be defined by the same formula (4), and $S^{(r)}$ is again a function of r only.

In the SS case (simultaneous, symmetric), as well as in the TS case (temporal, stationary) one can prove the following theorem:

$$S^{(r)} \leqslant S^{(s)} + S^{(t)} \tag{6}$$

with

$$r = s + t\,. \tag{6'}$$

From (6) follows that

$$S^{(r)} \leqslant \sum_{j=1}^{m} S^{(t_j)} \tag{7}$$

with

$$\sum_{j=1}^{m} t_j = r\,. \tag{8}$$

Equality in (6) holds (in the TS case) if and only if

$$\left\{\begin{array}{l} p^{(r)}(x_1, \ldots, x_r) = p^{(s)}(x_1, \ldots, x_s)p^{(t)}(x_{s+1}, \ldots, x_r) \\ \text{or} \\ p^{(r)}(x_1, \ldots, x_r) = p^{(t)}(x_1, \ldots, x_t)p^{(s)}(x_{t+1}, \ldots, x_r)\,, \end{array}\right. \tag{9}$$

for all values of the x's. In the SS case, due to the invariance for permutation, conditions (9) are only two representatives of $\binom{r}{s}$ similar conditions.

Among the possible decompositions of the type (7), the case, $t_i = 1$ ($i = 1, 2, \ldots, m = r$) gives the maximum values to the expression given on the right side of (7). Thus

$$S^{(r)} \leqslant \sum_{i=1}^{m} S^{(t_i)} \leqslant rS^{(1)}\,. \tag{10}$$

If and only if

$$p^{(r)}(x_1, \ldots, x_r) = p^{(1)}(x_1)p^{(1)}(x_2) \ldots p^{(1)}(x_r)\,, \tag{11}$$

i.e., if and only if the r variables are completely independent, then $S^{(r)} = rS^{(1)}$. Therefore, it is quite natural to consider

$$C^{(r)} \equiv rS^{(1)} - S^{(r)} \geqslant 0 \tag{12}$$

as the total correlation that exists in a configuration of r variables, in both SS and TS cases. In particular

$$C^{(n)} \equiv nS^{(1)} - S^{(n)} \geqslant 0 \tag{13}$$

is the total correlation existing in the entire system of n variables. The problem is to decompose $C^{(n)}$ into non-negative terms $T^{(r)}$ $(r = 2, 3, \ldots, n)$ each representing contribution to $C^{(n)}$ from that portion of correlation which can be attributed characteristically to an r-variable configuration.

Taking a set of r variables, let us suppose that we observe a subset of $(r-1)$ variables and the last r-th variable separately, and we process the observed data for these two groups separately. Under these conditions, we shall recognize only the correlation $C^{(r-1)}$. Then, it will be natural to consider $C^{(r)} - C^{(r-1)}$ as the correlation characteristic to the r-variable configuration over and beyond the correlation existing among $(r-1)$ variables. Thus, we shall call

$$U^{(r)} = C^{(r)} - C^{(r-1)} = S^{(1)} + S^{(r-1)} - S^{(r)} \tag{14}$$

the first correlation index of range r. In virtue of (6), we are guaranteed that

$$U^{(r)} \geqslant 0\,. \tag{15}$$

We can see easily that

$$C^{(n)} = \sum_{r=2}^{n} U^{(r)} \tag{16}$$

which satisfies the conditions required of $T^{(r)}$ at the beginning.

It is important to see what $U^{(r)} = 0$ means. In the TS case this happens if

$$p^{(r)}(x_1, \ldots, x_r) = p^{(r-1)}(x_1, \ldots, x_{r-1})p^{(1)}(x_r) \tag{17}$$

for all values of the x's, or if

$$p^{(r)}(x_1, \ldots, x_r) = p^{(1)}(x_1)p^{(r-1)}(x_2, \ldots, x_r) \tag{18}$$

for all values of the x's, and otherwise not. (17) and (18) mean

$$p(x_r \,|\, x_1, x_2, \ldots, x_{r-1}) = p^{(1)}(x_r)\,, \tag{17$'$}$$

$$p(x_2, x_3, \ldots, x_r \,|\, x_1) = p^{(r-1)}(x_2, x_3, \ldots, x_r)\,. \tag{18$'$}$$

where the left hand sides are conditional probabilities. These equalities would not hold even if the series were a simple Markoff chain. This can be most readily seen by summing over $x_1, \ldots, x_{r-1}$ in (17), or over $x_3, \ldots, x_r$ in (18), which shows that if $U^{(r)} = 0$, $r > 2$, then $U^{(2)} = 0$ or $p^{(2)}(x_1, x_2) = p^{(1)}(x_1)p^{(2)}(x_2)$. That means that even in a simple Markoff chain, one will get $U^{(r)} \neq 0$ for $r > 2$. This is not a desirable situation, if one wants to interpret $U^{(r)}$ as the correlation over and above $(r-1)$ variable correlation in the TS case.

In the SS case, (17) implies the $\binom{r}{1}$ conditions which can be obtained from (17) by permutations of variables. In this case too, $U^{(r)}=0$ implies $U^{(r-1)}=U^{(r-2)}=\ldots=U^{(2)}=0$. However, one cannot condemn (16) simply for this reason. Indeed in this case, there is no concept of distance between variables, therefore the presence of two-variable correlation may very well imply automatically the presence of higher-range correlation. In general, a non-vanishing $C^{(n)}$ implies that not all U's can vanish. This in turn means that there is a certain value of r, beyond which all $U^{(r)}$ are finite.

The $U^{(r)}$ introduced above is the first difference of $C^{(r)}$. The second difference of $C^{(r)}$ has also an important meaning:

$$(19)\qquad W^{(r)}\equiv U^{(r)}-U^{(r-1)}=C^{(r)}-2C^{(r-1)}+C^{(r-2)}=-S^{(r-2)}+2S^{(r-1)}-S^{(r)}\,.$$

which will be called second correlation index of range r.

It is easy to show that

$$(20)\qquad W^{(r)}\geqslant 0\,.$$

In the TS case the equality in (20) holds if and only if

$$(21)\qquad p^{(r)}(x_1,\ldots,x_r)=\frac{p^{(r-1)}(x_1,\ldots,x_{r-1})\,p^{(r-1)}(x_2,\ldots,x_r)}{p^{(r-2)}(x_2,\ldots,x_{r-2})}\,.$$

for all values of the x's. Condition (21) can also be rewritten as

$$(22)\qquad p(x_r\,|\,x_1,\ldots,x_{r-1})=p(x_r\,|\,x_2,\ldots,x_{r-2}).$$

This means that the conditional probability of $y_r=x_r$ is not influenced by y_1. This can be adequately interpreted as absence of correlation of range r over and above the correlations that may exist within a sequence of length $(r-1)$. Thus the W's are useful quantities for the TS case.

In the SS case, $W^{(r)}=0$ is equivalent to (21) and to all the equations obtained from (21) by permutation of variables.

It is easy to see that we can use $W^{(r)}$ as $T^{(r)}$:

$$(23)\qquad C^{(n)}=\sum_{r=2}^{n}(n-r+1)W^{(r)}\,,$$

with $S^{(0)}=0$, this is also a satisfactory development of $C^{(n)}$. See, S. WATANABE: *IRE Profess. Group. Inform. Theor. Symposium* 1954, p. 85.

It is conceivable to use still higher order differences of $C^{(r)}$ for the expansion of $C^{(n)}$. A particularly interesting form is

$$(24)\qquad C^{(n)}=\sum_{r=2}^{n}\binom{n}{r}F^{(r)}\,,$$

where $F^{(r)}$ is the r-th difference of $C^{(r)}$, *i.e.*,

$$(25)\qquad F^{(r)}=\left[-\binom{r}{r}S^{(r)}+\binom{r}{r-1}S^{(r-1)}-\binom{r}{r-2}S^{(r-2)}+\ldots+(-1)^r\binom{r}{1}S^{(1)}\right].$$

However, in the first place, $F^{(r)}$ can be both positive and negative. In the second place, the condition $F^{(r)}=0$ (which should mean the absence of r-variable correlation) cannot be clearly interpreted in terms of probabilities. In the SS case with $g=2$, $F^{(3)}$ calculated for $p(111)=\frac{1}{2}$, $p(000)=\frac{1}{2}$ becomes -1. Thus, expansion (24) is not satisfactory.

Finally, it should be noted that the well-known expression of $p^{(n)}(x_1, x_2, \ldots, x_n)$ in terms of conditional probabilities:

$$(26)\qquad p^{(n)}(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1x_2)\ldots p(x_n|x_1, \ldots, x_{n-1})$$

leads to

$$(27)\qquad S^{(n)} = S^{(1)} + (S^{(2)} - S^{(1)}) + \ldots + (S^{(n)} - S^{(n-1)}),$$

where each term $(S^{(r)} - S^{(r-1)})$ has no clear meaning as far as correlation is concerned.

As conclusion, it can be said that in the TS case the expansion (23) in terms of $W^{(r)}$ is definitively the most appropriate. In the SS case, the expansion (16) in terms of $U^{(r)}$ has a clear meaning in the following operational conditions. When n variables are given the observer starts with examining the information content of only one variable, and then of two variables, and then of three variables, etc..., at each stage increasing one more variable. When he passes from r variables to $(r+1)$ variables, if the last one were completely independent from the other r variables then the information content would be $S^{(r)}+S^{(1)}$. But when he observes all the $(r+1)$ variables together, he discovers that the information content is only $S^{(r+1)}$. The difference $S^{(r)}+S^{(1)}-S^{(r+1)}$, he would call the additional correlation $U^{(r+1)}$ peculiar to the $(r+1)$ variables.

In a similar manner, the quantity $(S^{(s)}+S^{(t)}-S^{(s+t)})$ has the following clear meaning. The observer first observes a group of s variables and a group of t variables separately. The total information is $S^{(s)}+S^{(t)}$. Next he observes the $(s+t)$ variables together. Then he gets information $S^{(s+t)}$. The decrease in information $(S^{(s)}+S^{(t)}-S^{(s+t)})$ is to be considered as the correlation existing between the group of s variables and the group of t variables.

The domains of application of the present results are numerous ranging from physics to psychology, sociology, etc., just to name a few. In physics, an n-fermion system has to obey the Pauli exclusion principle. This principle is a correlation in the present sense, since if a particle occupies a certain quantum state, then other particles are prevented from occupying this quantum state. The two-body correlation $C^{(2)}$ due to this effect can be estimated to be $\log[2n/(n-1)]$.

Dr. Nancy Anderson and Mr. John Ross of IBM Research Laboratory are investigating the relationship between the correlation in the sense of this report and the correlation coefficients in the usual sense, having in mind applications to psychology.

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3º Trimestre

On the Detection of Gaussian Signals in Gaussian Noise.

D. Slepian

Bell Telephone Laboratory - Murray Hill, N.J.

1. – Introduction.

The problem of detecting a Gaussian signal in Gaussian noise has been discussed by a number of authors during the past decade. A recent paper by Middleton [1] on this subject contains references to much of the earlier work. Here we comment on several aspects of this problem generally overlooked in the past.

The problem treated can be stated as follows. An observer has available to him a sample of finite duration, $x(t)$, $0 \leqslant t \leqslant T$, of a stationary Gaussian process. It is known to him that $x(t)$ is either a sample from the Gaussian ensemble A with mean zero and power density spectrum $\varphi_A(f)$ or a sample from the Gaussian ensemble B with mean zero and power density spectrum $\varphi_B(f)$. The observer is to decide whether $x(t)$, $0 \leqslant t \leqslant T$ came from A or B. In most engineering applications, A is interpreted as signal plus noise and B as noise alone.

The observer's decision as to which ensemble $x(t)$ came from can be in error in two different ways; he can assert that $x(t)$, came from A when indeed it came from B; or he can assert that $x(t)$, came from B when indeed it came from A. We denote the probabilities of these two types of error by p_A and p_B respectively.

The main result that has been obtained by the author is that for the spectra generally considered in engineering problems, it is possible for the observer to make his decision with vanishingly small probability of error of either kind.

Here we report on one part of our findings. The result described below can be generalized and extended. For full details the reader is referred to the complete paper to appear in *IRE Trans. Profess. Group Inform. Theor.*, June 1958.

2. – Properties of a certain quadratic form.

From the observed sample $x(t)$, $0 \leqslant t \leqslant T$ form the quadratic form:

$$(1) \qquad y_n = \left(\frac{n}{T}\right)^{2m-2} \sum_{j=0}^{q-1} \left(\sum_{h=0}^{m} \binom{m}{h} (-1)^h x\left\{\frac{(jm+h)T}{n}\right\}\right)^2 ,$$

where $n = mq$, m and q are positive integers. This form uses only the values $x(jT/n)$, $j = 0, 1, \ldots, n$. We first investigate some properties of y_n when it assumed that $x(t)$ is a stationary Gaussian process with mean zero and covariance function $r(\tau) = Ex(t)x(t+\tau)$. We shall further assume that the spectrum of the process

$$\varphi(f) = \int_{-\infty}^{\infty} \exp\left[2\pi i f \tau\right] r(\tau) \mathrm{d}\tau ,$$

has the asymptotic behaviour for large f

$$\varphi(f) \sim \frac{a}{f^{2s}} ,$$

where s is a positive integer.

The expected value of y_n is

$$Ey_n = \frac{n^{2m-1}}{mT^{2m-2}} \sum_{k=0}^{m} \sum_{l=0}^{m} \binom{m}{k} \binom{m}{l} (-1)^{k+l} r\left[\frac{(l-k)T}{n}\right] .$$

On introducing the Fourier integral representation for r, this can be written

$$Ey_n = \frac{n^{2m-1}}{mT^{2m-2}} \int_{-\infty}^{\infty} \mathrm{d}f \varphi(f) \sum_{k=0}^{m} \sum_{l=0}^{m} \binom{m}{k} \binom{m}{l} (-1)^{k+l} \exp\left[2\pi i (l-k) f \,(T/n)\right] =$$

$$= \frac{2^{2m} n^{2m-1}}{mT^{2m-2}} \int_{-\infty}^{\infty} \mathrm{d}f \varphi(f) \sin^{2m} \frac{\pi T f}{n} = \frac{2^{2m} \pi^{2m-1} T}{m} \int_{-\infty}^{\infty} \mathrm{d}\xi \left(\frac{n\xi}{\pi T}\right)^{2m} \varphi\left(\frac{n\xi}{\pi T}\right) \left(\frac{\sin \xi}{\xi}\right)^{2m} .$$

It then follows easily from the asymptotic behavior of φ that if $s > m$,

$$\lim_{q\to\infty} Ey_{qm} = 0 ,$$

while, if $s = m$,

$$\lim_{q\to\infty} Ey_{qm} = ac_m ,$$

where

$$c_m = \frac{2^{2m} \pi^{2m-1} T}{m} \int_{-\infty}^{\infty} \mathrm{d}\xi \left(\frac{\sin \xi}{\xi}\right)^{2m} .$$

The variance of y_n can also be computed in a straightforward manner. One finds

$$\operatorname{Var} y_n = \left(\frac{n}{T}\right)^{4m-4} \sum_{j,\sigma=0}^{q-1} \sum_{\substack{k,l\\ \mu,\nu=0}}^{m} \binom{m}{k}\binom{m}{l}\binom{m}{\mu}\binom{m}{\nu} (-1)^{k+l+\mu+\nu},$$

$$\Big[r\left\{[(j-\sigma)m+k-\mu]\frac{T}{n}\right\} r\left\{[(j-\sigma)m+l-\nu]\frac{T}{n}\right\} + \\ + r\left\{[(j-\sigma)m+k-\nu]\frac{T}{n}\right\} r\left\{[(j-\sigma)m+l-\mu]\frac{T}{n}\right\}\Big].$$

Introduce the Fourier integral representation of r, interchange summation and integration and perform the sums. There results:

$$\operatorname{Var} y_n = 2^{4m-1}\left(\frac{n}{T}\right)^{4m-4} \int_{-\infty}^{\infty} \mathrm{d}f \int_{-\infty}^{\infty} \mathrm{d}f' \varphi(f)\varphi(f') \cdot \\ \cdot \left[\frac{\sin \pi m q(f-f')(T/n)}{\sin \pi m(f-f')(T/n)}\right]^2 \left[\sin \pi f \frac{T}{n} \sin \pi f' \frac{T}{n}\right]^{2m},$$

or, after an appropriate change of variables,

$$(2) \qquad \operatorname{Var} y_n = \frac{2^{4m-1}\pi^{2m-2}T^2}{n^2} \int_{-\infty}^{\infty} \mathrm{d}\xi \int_{-\infty}^{\infty} \mathrm{d}\eta \left(\frac{n\xi}{\pi T}\right)^{2m} \varphi\left(\frac{n\xi}{\pi T}\right) \left(\frac{n\eta}{\pi T}\right)^{2m} \varphi\left(\frac{n\eta}{\pi T}\right) \cdot \\ \cdot \left[\frac{\sin\xi}{\xi}\, \frac{\sin\eta}{\eta}\right]^{2m} \left[\frac{\sin mq(\xi-\eta)}{\sin m(\xi-\eta)}\right]^2.$$

We now show that $\lim_{q\to\infty} \operatorname{Var} y_{mq} = 0$. Since all factors in the integrand of (2) are non-negative and since $f^{2m}\varphi(f)$ is bounded from above by our assumption, it follows that

$$(3) \qquad \operatorname{Var} y_{mq} < \frac{d}{q} h_q,$$

where d does not depend on n or q and

$$h_q = \frac{1}{q} \int_{-\infty}^{\infty} \mathrm{d}\xi \int_{-\infty}^{\infty} \mathrm{d}\eta \left[\frac{\sin(\xi/m)}{\xi/m}\, \frac{\sin(\eta/m)}{\eta/m}\right]^{2m} \left[\frac{\sin q(\xi-\eta)}{\sin(\xi-\eta)}\right]^2.$$

Now

$$(4) \qquad h_q = \frac{1}{q} \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \int_{j\pi}^{(j+1)\pi} \mathrm{d}\xi_j \int_{k\pi}^{(k+1)\pi} \mathrm{d}\eta_k \left[\frac{\sin(\xi_j/m)}{\xi_j/m}\, \frac{\sin(\eta_k/m)}{\eta_k/m}\right]^{2m} \left[\frac{\sin q(\xi_j-\eta_k)}{\sin(\xi_j-\eta_k)}\right]^2 = \\ = \frac{1}{q} \int_0^{\pi} \mathrm{d}\xi \int_0^{\pi} \mathrm{d}\eta \left[\frac{\sin q(\xi-\eta)}{\sin(\xi-\eta)}\right]^2 h(\xi)\, h(\eta) = \frac{1}{q} \int_0^{\pi} \mathrm{d}x \left[\frac{\sin qx}{\sin x}\right]^2 g(x),$$

where

$$h(\xi) = \sum_{j=-\infty}^{\infty} \left[\frac{\sin (1/m)(\xi + j\pi)}{(1/m)(\xi + j\pi)}\right]^{2m},$$

and

$$g(x) = \int_{x}^{2\pi - x} \mathrm{d}y h \left[\frac{1}{2}(y + x)\right] h \left[\frac{1}{2}(y - x)\right].$$

The square bracket in the last integral of (4) is the Fejer kernel studied in Fourier theory [2]. As $q \to \infty$, $h_q \to \pi g(0+)$ which is finite. From (3) it follows that Var $y_{mq} \to 0$ as $q \to \infty$.

3. – The detection problem.

The preceding paragraphs show that if $\varphi \sim a/f^{2s}$, then y_{mq} of (1) converges in probability to zero if $s > m$, and to a c_m if $s = m$. Consider now the problem of distinguishing between the stationary Gaussian ensembles A and B where $\varphi_A \sim a/f^{2m}$ and $\varphi_B \sim b/f^{2(m+p)}$, $p \geqslant 0$. Form the test function (1) from the observed samples $x(jT/n)$, $j = 0, 1, ..., n$, $n = mq$. If $p > 0$, choose a threshold $\boldsymbol{y}$ between zero and $c_m a$. Use the decision rule « sample came from A if $y_{qm} \geqslant \boldsymbol{y}$; sample came from B if $y_{qm} < \boldsymbol{y}$. » If $p = 0$ and $a > b$, choose a threshold $\boldsymbol{y}$ between $c_m a$ and $c_m b$. Use the decision rule « sample came from A if $y_{qm} \geqslant \boldsymbol{y}$; sample came from B if $y_{qm} < \boldsymbol{y}$. » If $p = 0$ and $a < b$, reverse the decision rule. By choosing q, and hence n, sufficiently large p_A and p_B can be made arbitrarily small.

It is to be noted the foregoing results say nothing about the case in which φ_A and φ_B have identical asymptotic behavior. It is not known in this case whether or not sequences of tests based on $x(t)$, $0 \leqslant t \leqslant T$ can be constructed which make p_A and p_B arbitrarily small. This case is often met in engineering applications in which A is interpreted as signal plus noise, B as noise alone and it is assumed that the signal spectrum falls off more rapidly with increasing f than the noise spectrum.

Whether or not perfect detection is possible in this case, the results presented show this model to be a poor one for the engineering problem. For, one can always suppose added to the signal spectrum an arbitrarily small amount of Gaussian signal power at extremely high frequencies so that indeed the asymptotic behaviours of φ_A and φ_B are different and perfect detection is possible. The addition of this arbitrarily small amount of signal power at arbitrarily high frequencies is certainly non-physical. Thus by altering non-physically significant parameters in the detection model, we can be assured of perfect detection.

The most obvious short-coming of this model is the assumption that the observer has perfect knowledge of φ_A and φ_B beforehand. The author feels that proper inclusion of the observer's uncertainty about φ_A and φ_B will preclude the possibility of perfect detection on the basis of an observation of a sample of finite duration, and will result in a model that more accurately describes the situation actually encountered in engineering application. He looks forward to the development of such a model.

REFERENCES

[1] D. Middleton: *On the Detection of Stochastic Signals in Additive Normal Noise,* Part I, *IRE Trans.*, Vol. IT-3, 86 (June 1957).

[2] E. T. Whittaker and G. N. Watson: *Modern Analysis* (New York, 1947), p. 170.

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3º Trimestre

On the Realizability Problem for Irredundant Boolean Networks.

L. Löfgren

Swedish Research Institute of National Defence - Stockholm

1. – Introduction.

We will give a closed theory of a certain non trivial class of Boolean networks, the irredundant networks.

Two criteria will be derived by which we can decide if a Boolean function has an irredundant network or not. Each is necessary and both together are sufficient. The two criteria are called the *c*-criterion and the subrearrangement Γ_L-criterion.

It is possible to refer the general minimization problem for Boolean networks to a series of existence problems of the kind treated here. We will restrict the following treatment only to irredundant networks.

The particular reason for this study of irredundant networks was a more general study of networks with a minimum redundancy-number of branches for a certain protection against branch errors.

We mean with a Boolean function a disjunction of clauses. A clause is a conjunction of literals (where no letter appears twice). A literal is a letter either affirmed or negated. Only the values 0 or 1 can be assigned to a letter. Each Boolean function can be transformed into a standard form, for which we choose an indispensable prime implicant form (for prime implicants, see Quine [3]).

We will always refer an irredundant network to a Boolean function of such a standard form. We say by definition that a network is irredundant if there is a 1 : 1 correspondence between its branches and the different literals of the «corresponding» Boolean function of standard form (the literals of a standard form are irredundant). After having derived the *c*-criterion, we will give a more precise meaning of the word «corresponding» (Sect. **3**).

Giving the sub-rearrangement criteria of Sect. **4**, we also give the solution to the following topological problem: How can a matrix of incidence (with elements reduced mod 2) which corresponds to a linear graph be characterized. (Or: which matrices can be reduced so as to contain only two 1's in each column.) This problem has been put forth by OKADA [2] and by SESHU [4], but to the author's knowledge no solution has been published so far.

In the following we will be concerned only with 2-terminal networks. The extension of the treatment to n-terminal networks is quite straightforward (will be published elsewhere).

2. – The Veblen incidence matrix H and the loop matrix L.

Let us consider a graph of a Boolean network with c branches and r vertices. It is uniquely represented (except for row permutations) with the incidence matrix

$$H = \begin{array}{c|cccccc|}
 & a & b & c & d & \dots & \\
1 & & & & & & \\
2 & & & & & & \\
3 & & & & & & \\
\cdot & & & & & & \\
\cdot & & & & & & \\
\cdot & & & & & &
\end{array} = \|\eta_{ij}\| , \tag{1}$$

with c columns corresponding to the branches $(a, b, c, d, \dots)$ and r rows corresponding to the vertices $(1, 2, 3, \dots)$ so that η_{ij} is 1 if the i-th vertex is incident with the j-th branch, and $\eta_{ij} = 0$ otherwise. Thus the i-th row of H is the symbol for the set of branches which are incident with the vertex i. The j-th column is the symbol for the point pair incident with the branch j. Since each branch is incident with exactly two points, every column must contain two 1's. Conversely, any matrix whos elements are 0's and 1's and which is such that each column contains exactly two 1's and each row at least one 1, can be regarded as the incidence matrix of a branch-vertex graph (not necessarily connected, however).

Let us denote a path between two terminal vertices with the row matrix

$$\begin{array}{cccccc}
 & a & b & c & d & \dots \\
P = \| & 0 & 1 & 1 & 0 & \dots \|
\end{array} \tag{2}$$

with 1's in those columns which correspond to branches of the path and with 0's in the other columns. The matrix product $H \cdot \overline{P}$ (with addition mod 2 and ordinary multiplication), where $\overline{P}$ is the transpose of P, will then be a column matrix with only two 1's, one for each of the two terminal vertices, and with 0's in all other rows. This because the i-th row of H (the symbol for the set of branches of the graph incident with point i) will produce a 1 in the product if and only if there is an odd number of branches contained in P which are incident with the point i. This is a general criterion of an end point in a path (which also may contain loops).

Let us then with P represent all paths between two terminal vertices of a graph (one row for each path). The product $H \cdot \bar{P}$ will then have two strings of 1's in the rows which correspond to the terminal vertices (and 0's elsewhere). By adjoining to H and to P an extra terminal branch T between the two terminal vertices, the path matrix becomes a loop matrix L_T and H will be denoted H_T^2 (the index 2 refers to the fact that each column has exactly two 1's). The matrix product

$$H_T^2 \cdot \bar{L}_T = 0 \tag{3}$$

is now a matrix in which every element is a 0.

Let us now consider a Boolean function of a standard form, for instance

$$B = abc \cup a'b' \cup \ldots,$$

and let a branch, say b', corresponding to the literal b', have the value 1 if the branch (switching element) is transmitting (« make ») and the value 0 if the branch is non-transmitting (« break »). Thus « make » and « break » between the end points of two series-connected branches a' and b' is represented by $a' \cdot b'$, and for a parallel connection the value $a' \cup b'$ is obtained. (Compare SHANNON [5,6]). So a Boolean function uniquely specifies a loop matrix L_T in which the rows correspond to the clauses of B with a 1 in the T-column, and 1's in all those columns which as literals constitute the clause (and 0's in the other columns).

So in order to investigate the existence of an irredundant network for B we have to solve equation (3) for H_T^2, when L_T is known.

Let us consider the complete solution H_T of (3). It is evidently a group G_H under addition mod 2, for if α and β are two solution elements, so is $\alpha+\beta$ mod 2. Further the identity element of G_H is a string of 0's (which obviously also is a solution of (3)). The inverse of an element of G_H evidently is the element itself. Finally since the operation of addition mod 2 (digit by digit) is commutative, G_H is an abelian group.

If we instead consider H_T^2 of (3) known and solve for L_T, the complete solution G_L is also an abelian group under addition mod 2.

3. – First kind of realizability criterion (the c-criterion).

When determining G_H from (3) with L_T specified, we must require that the part of G_H which actually forms the H_T^2-solution must be such that we from this part (and (3)) can determine an L_T' which is in acceptable agreement with the specified L_T. We mean with an acceptable agreement between L_T' and L_T that the corresponding sets B' and B satisfy

$$B \subset B' \subset B_r \,, \tag{5}$$

B_r is a complete redundant form of B, *i.e.* B_r contains every element that implies B (beside the B-elements, also elements which subsume B-elements and other dispensable clauses) and vanishing elements containing a letter both affirmed and negated.

Let us denote with Γ_L a generator set of G_L. Let us further transform Γ_L under element addition (mod 2) so as to contain a single element γ with a 1 in the T-position and a rest, the set Γ. Each element of Γ has a 0 in the T-position. This division of Γ_L into γ and Γ,

$$\Gamma_L = \gamma \cup \Gamma \,, \tag{6}$$

is always possible since Γ_L is a generator set. In order to obtain L_T' from G_L we divide G_L into a subgroup g generated by Γ and a coset c:

$$c = \gamma + g \pmod 2 \,. \tag{7}$$

Evidently all elements of the coset c have a 1 in the T-position and thus

$$c = L_T' \,. \tag{8}$$

Evidently c covers L_T and from (5) we obtain the following necessary criterion:

c-criterion: A necessary condition for a Boolean function B of standard form (specifying L_T) to have an irredundant network is that the coset c (7) of the complete group generated by L_T is contained in the complete redundant form B_{rT} of B:

$$B_{rT} \supset c = \gamma + g \quad \pmod 2 \,. \tag{9}$$

(It should be observed that c contains multiple-loops through T. These are not dismissed from the criterion because a corresponding clause in B always subsumes a clause which corresponds to a single loop through T.)

Evidently any complete independent set of L_T can be chosen for the generator set Γ_L (6). The coset c is still the same.

We mean with networks « corresponding » to a Boolean function B, all those networks (H^2_T) which have the same c (determined from B according to the above) and which can be determined from (3). In the next section **4** we will see how to decide if any such networks exist, and if so, how to determine all of them.

4. – Second type of realizability criterion (the subrearrangement criterion).

Evidently a solution H^2_T of (3) must contain so many elements of G_H so as to determine G_L from the same equation (3). This implies that H^2_T must contain at least a generator set Γ_H of G_H. Furthermore we must require that each column of H^2_T contains precisely two 1's. This implies a relation between the elements of H^2_T. Since all (say $r-1$) elements of a generator set are independent H^2_T must at least contain r elements, the r-th being the sum (mod 2) of the $r-1$ generator elements.

Lemma 1. If an irredundant H^2_T-solution of (3) (with L_T specified by Γ_L) exists, it must contain one of the generator sets Γ_H (each having $r-1$ elements) of G_H and one further element (the r-th element) being the sum mod 2 of the elements of Γ_H.

The *proof* will run as follows. Suppose that H^2_T consists of a generator set (with elements $\gamma_1, \gamma_2, \ldots, \gamma_{r-1}$) and of the $k+1$ depending elements $\gamma_r, \gamma_{r+1}, \ldots, \gamma_{r+k}$. The corresponding $k+1$ relations may be written:

$$(10)\quad \left|\begin{array}{llllll} \Sigma_0\gamma_i = 0 & \text{(mod 2). The set of elements covered by} & \Sigma_0 & \text{is denoted} & S_0 \\ \Sigma_1\gamma_i = 0 & \text{»} \qquad\qquad\qquad \text{»} & \Sigma_1 & \text{»} & S_1 \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ \Sigma_k\gamma_i = 0 & \text{»} \qquad\qquad\qquad \text{»} & \Sigma_k & \text{»} & S_k \end{array}\right.$$

It is possible to show that we can always write the relations (10) in such a form that all the sets S_i are disjoint (and so that their union covers all the $r+k$ elements). (Compare also VEBLEN [7]). Recalling that an element of H^2_T represents the branches which are incident with the corresponding vertex, this means that each branch which has one vertex in the set S_i must also have the other vertex in S_i. Since the sets S_i are disjoint, the graph must consist of $k+1$ disconnected subgraphs. But all the branches of a disconnected subgraph which does not contain the T-element must be redundant. For the clauses of B, *i.e.* the paths through T, are only determined by the branches of the subgraph which contains the T-element. Hence an H^2_T-solution (an irredundant network) cannot contain disconnected subgraphs and k must be zero.

The question is thus referred to: which Γ_H-sets of G_H are Γ_{H^2}-sets, *i.e.* contain at most two 1's in each column? (By adjoining an r-th element according to lemma 1, each column will then have precisely two 1's.)

There are

$$(11)\qquad N(r) = (2^{r-1}-2^0)(2^{r-1}-2^1)(2^{r-1}-2^2)\dots(2^{r-1}-2^{r-2})/(r-1)!$$

different generator sets corresponding to a group generated by $r-1$ elements. For example $N(7)=28\,901\,376$.

It is useful to observe that any generator set of G_H can be obtained from a specific generator set Γ_H only by repeated operations of the type: replace one row by the sum (mod 2) of this row and some other row (a row-replacement operation).

Generally it is quite easy to convert a Γ_H-matrix into Γ_{H^2}-form by row-replacement operations, *i.e.*, if a Γ_{H^2}-form exists. A strategy in the beginning of the operations is to add two rows which have many 1's in the same columns together, and to replace one of them by the sum so as to obtain an increase in the number of 1's in a number of columns as small as possible.

However if no Γ_{H^2}-form exists, we must know exactly that none of all the $N(r)$, see (11), Γ_H-sets is of Γ_{H^2}-form. Such an information is nicely obtained from the subrearrangement criterion to be derived in what follows. Also if a Γ_{H^2}-form exists but has not been found by the row-replacement procedure, it is obtained from the subrearrangement criterion, which in addition gives all H^2_T-forms that correspond to the coset c, if more than one exist.

The idea behind the subrearrangement criterion is to derive a subproblem whose solutions will give character to the main realizability problem. We will prove that it is possible to give the subproblem essentially the same nature as that of the main problem, but with immediately obtained solutions, which restrict the number of investigations for the main problem to the many-valuedness of the subproblem.

An H^2_T-solution must be a non-separable graph, but for the subproblem we must admit for solutions, H^2, which may consist of separable loop-graphs. We will say that a subgraph is tied to another subgraph if it is connected to it at three or more vertices. If it is connected at only two vertices it is loosely tied, and if it is connected at one or no vertex, the graph is separable. A single branch with its two vertices will not be considered as a subgraph.

For the criterion we need the following lemma.

Lemma 2. A matrix of incidence H^2 (of r rows) of a loop-graph (a graph where each branch is incident with a loop) is defined by its $\underline{\Gamma}$-matrix which consists of $r-1$ arbitrarily chosen rows of H^2 (the r-th row is the sum of the rows of $\underline{\Gamma}$). The elements of $\underline{\Gamma}$ need not be independent. All possible proper row-replacements on $\underline{\Gamma}$ which maintain the H^2-form (two 1's in each column) are repeated applications of the following types. (A pure row-reordering row-replacement operation is considered improper, because the arrangement of the elements in an incidence matrix is without significance.)

i) If H^2 contains loosely tied subgraphs, say H_1 which has only the two vertices r_i and r_j in common with another subgraph, then the row-replacement

$$r_i \to r_i + \sum_{H_1}^{\text{exc } r_i,\, r_j} (r_\nu)\,,$$

$$r_j \to r_j + \sum_{H_1}^{\text{exc } r_i,\, r_j} (r_\nu)\,,$$

(r_i is replaced by the sum mod 2 of itself and all the vertices of H_1 except r_i and r_j, and similar for r_j) is possible for any loosely tied subgraphs of H^2. This operation corresponds geometrically to a change of connectivity of H_1.

ii) If $\underline{\Gamma}$ contains (non-zero) elements which depend on its other non-zero elements according to the equation system (10), and contains zero-elements $r_\nu(0)$, then any row-replacement

$$r_i \to r_i + \Sigma_j + r_\nu(0)$$

is possible, where r_i is any row of $\underline{\Gamma}$, Σ_j is any disjoint zero-sum of (10), and $r_\nu(0)$ is any zero-element.

iii) If H^2 is separable:

α) If disconnected subgraphs exist say H_k and H_l, incident with the vertices r_k and r_l respectively, these subgraphs can be joined at a single vertex (cut-vertex) r_{kl} (formed when r_k and r_l coalesce). The corresponding row-replacement,

$$r_k \to r_k + r_l = r_{kl}\,,$$

$$r_l \to \sum_{H_l} r_\nu = r(0) = 00 \ldots 00\,,$$

is possible for any pairs of disconnected subgraphs and for any of their vertices.

β) If a cut-vertex r_{ij} and a zero-row $r(0)$ exist, the cut-vertex can be split into r_i and r_j belonging to the generated disconnected subgraphs H_i and H_j. The corresponding row-replacement

$$r(0) \to r(0) + \sum_{H_i}^{\text{exc } r_{ij}} (r_\nu) = r_i$$

$$r_{ij} \to r_{ij} + r_i = r_j$$

is possible for any zero-row $r(0)$ and any cut-vertex r_{ij}.

γ) If a cut vertex r_{ij} exists, we can change it to r_{kl} (operation β and α) also without having a zero-row to disposal by the row-replacement

$$r_k \to r_k + r_l = r_{kl}$$

$$r_{ij} \to r_{ij} + \sum_{H_l}^{\text{exc } r_{ij}} (r_\nu) \quad = r_i$$

$$r_l \to r_l + \sum_{H_l}^{\text{exc } r_l,\, r_{ij}} (r_\nu) = r_j$$

H_l is the separable subgraph of H^2 which contains r_l but not r_k: When r_k and r_l coalesce, r_{kl} is formed. At the same time r_{ij} is split into r_i and r_j, where r_j but not r_i is contained in H_l after the change of cut-vertex.

Proof. The proof will run as follows. An H^2-matrix determines a graph uniquely and conversely, provided the matrix has only non-zero elements. H^2 specifies uniquely a complete set of loops (G_L), and all other matrices with the same G_L (or with loops in 1:1 correspondence with the loops of H^2) must be obtained by row-replacements on H^2 (a row-replacement on H^2 does not change G_L). All matrices with loops in 1:1 correspondence with the loops of H^2 are obtained by connetivity changes of H^2-subgraphs corresponding to (i and iii). This follows from a theorem of WHITNEY [9]:

« If there is a 1:1 correspondence between the branches of the two graphs G and G' so that loops correspond to loops, then the graphs are strictly 2-isomorphic. »

Two graphs G and G' are 2-isomorphic if one can be transformed into the other by the connectivity changes:

i) If $G = H_1 + H_2$, where H_1 and H_2 have just the vertices r_i and r_j in common and these vertices are connected in both H_1 and H_2, then H_1 is turned around at these vertices.

iii) Break a graph at a single vertex into two connected pieces, or join two connected pieces (not connected to each other) at a single vertex.

There is only one proper independent row-replacement for each change of connectivity in the H^2-graph if $\underline{\Gamma}$ is a generator set. These are easily found to be the ones listed above (i and iii). However if $\underline{\Gamma}$ has depending elements, the row-replacements of ii are possible. For let us represent a row-replacement by a square transformation matrix T of $r-1$ rows. Suppose a certain change of connectivity transforms the matrix $\underline{\Gamma}$ to $\underline{\Gamma}^*$:

$$T\underline{\Gamma} = \underline{\Gamma}^* . \tag{12}$$

Suppose there exists another matrix T' for this transformation. Then:

$$(T + T')\underline{\Gamma} = 0 \qquad (\text{mod } 2) . \tag{13}$$

If $\underline{\Gamma}$ is a generator set the only solution is $T' = T$. If however $\underline{\Gamma}$ has k dependent rows according to (10) and also contains n zero-rows $r_\nu(0)$, we first transform $\underline{\Gamma}$ so as to contain $k+n$ zero-rows and a proper generator set Γ (of $r-1-k-n$ rows). For this transformed $\underline{\Gamma}$ the solution of (13) is:

$$(14) \qquad T' = T + \left\| \begin{matrix} \overbrace{\ulcorner \qquad \ulcorner}^{k+n} & 0 & 0 & \cdots & 0 \\ & 0 & 0 & \cdots & 0 \\ \cdots & 0 & 0 & \cdots & 0 \\ & & & \vdots & \\ & 0 & 0 & \cdots & 0 \\ \llcorner \qquad \lrcorner & \underbrace{0 \quad 0 \quad \cdots \quad 0}_{r-1-k-n} \end{matrix} \right\|$$

where the indicated matrix has $k+n$ arbitrary columns and $(r-1-k-n)$ zero-columns. This means that we for any allowable row-replacement T also can replace any row by itself and any zero-rows $r_\nu(0)$ or any zero-sums Σ_j according to ii.

We are now able to prove the following constructive Γ_{H^2}-criterion.

The subrearrangement Γ_{H^2}-criterion. – A Γ_H-matrix is converted by row-replacement operations so as to contain at most two 1's and at least one 1 in ν' of its c columns and more than two 1's in the remaining ν'' columns. A corresponding Γ_L-matrix is transformed with row-replacement operations so that $(\Gamma_L)_{\nu''}$ (the ν''-columns of Γ_L) is converted into independent rows and zero-rows. The rows of the so transformed Γ_L-matrix which contain these independent rows of $(\Gamma_L)_{\nu''}$ are deleted, and the remaining rows are denoted Γ_L'. All non-empty columns, say $\underline{\nu}'$, of $(\Gamma_L')_{\nu'}$ are denoted $(\Gamma_L')_{\underline{\nu}'}$. The same columns of Γ_H are denoted $(\Gamma_H)_{\underline{\nu}'}$. The only row-replacements which may convert Γ_H into Γ_{H^2}-form are those which correspond to rearrangements of subgraphs according to lemma 2 with $\underline{\Gamma} = (\Gamma_H)_{\underline{\nu}'}$. If the r-th row of $(H^2)_{\underline{\nu}'}$, which is not present in $(\Gamma_H)_{\underline{\nu}'}$, is involved in a row-replacement, a row r_i of $(\Gamma_H)_{\underline{\nu}'}$, which is not involved, is replaced by the r-th. (The ν'-ν''-division is unaffected by this replacement). Should there be an empty column of Γ_H, no Γ_{H^2}-form of Γ_H exists.

Proof. Let us first consider the case when there is an empty column of Γ_H. Then any other form of Γ_H must also have only zeroes in this column, and we know from lemma 1 that no H_T^2-form exists. Let us then consider the division of the columns into a $\underline{\nu}'$ and a $\underline{\nu}''$ $(= c - \underline{\nu}')$ part. We will prove that $(\Gamma_H)_{\underline{\nu}'}$, as defined in the criterion is a $\underline{\Gamma}$-set (compare lemma 2) of an H^2-matrix corresponding to a loop-graph with all its loops in 1:1 correspondence with all the loops of H_T^2 which are only incident with the $\underline{\nu}'$-branches. Since $(\Gamma_H)_{\underline{\nu}'}$

has exactly as many rows as Γ_H, and since a row-replacement on $(\Gamma_H)_{\underline{\nu'}}$ operates in the same way on Γ_H, it follows that the row-replacements according to the criterion are the only ones which can possibly convert Γ_H into Γ_{H^2}-form. Other row-replacements on $(\Gamma_H)_{\nu'}$ which conserves the $(\Gamma_{H^2})_{\underline{\nu'}}$-form can be expressed as further improper replacements, and an improper replacement cannot diminish the number of 1's in a ν''-column to two or one. Let us now denote with G'_L the group (with the zero-row deleted) generated by Γ'_L and the rest of the group G_L with G''_L (with the zero-row deleted). The $\underline{\nu}'$-columns of G'_L and G''_L are denoted with $(G'_L)_{\underline{\nu'}}$ and $(G''_L)_{\underline{\nu'}}$. The remaining columns of G'_L and G''_L are denoted with $(G'_L)_{\underline{\nu''}}$ and $(G''_L)_{\underline{\nu''}}$ respectively. Equation (3) is with this division equivalent to

$$\Gamma_H \cdot \overline{G}'_L = 0\,, \tag{15}$$

$$\Gamma_H \cdot \overline{G}''_L = 0\,. \tag{16}$$

Since $(G'_L)_{\underline{\nu''}}$ has only zero-elements, (15) implies

$$(\Gamma_H)_{\underline{\nu'}} \cdot (\overline{G}'_L)_{\underline{\nu'}} = 0 \tag{17}$$

and from (16) we obtain

$$(\Gamma_H)_{\underline{\nu''}}\overline{G} \cdot ({}''_L)_{\underline{\nu''}} = (\Gamma_H)_{\underline{\nu'}} \cdot (\overline{G}''_L)_{\underline{\nu'}}\ . \tag{18}$$

Hence when we want to determine Γ_H from (3) this is equivalent to a determination of the $(\Gamma_H)_{\underline{\nu'}}$-part from (17) only, and then the remaining $(\Gamma_H)_{\underline{\nu''}}$-part is determined from (18). Hence we have referred the determination of $(\Gamma_H)_{\underline{\nu'}}$ to a problem of the same character as the problem of determining H^2_T from (3) with a specified group G_L. We know that $(G'_L)_{\underline{\nu'}}$, is a proper loop-group, because it consists of all those loops of G_L which are only incident with the $\underline{\nu}'$-branches, and it contains no zero-column. Hence the graph corresponding to $(\Gamma_H)_{\underline{\nu'}}$ (obtained by adjoining an r-th row) must be a loop-graph, not necessarily connected however since $(\Gamma_H)_{\nu'}$ may contain dependent rows and zero-rows. $(\Gamma_H)_{\underline{\nu'}}$ is, by definition, of $(\Gamma_{H^2})_{\underline{\nu'}}$-form and it follows from lemma 2 $((\Gamma_H)_{\underline{\nu'}} = \underline{\Gamma})$ that the sub-rearrangement criterion is correct.

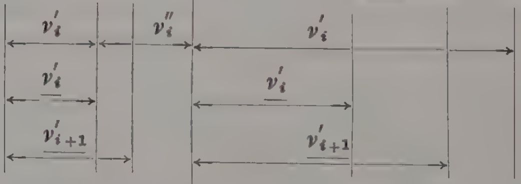
When trying to transform a Γ_H-matrix into Γ_{H^2}-form, this criterion can always be applied and is an effective short cut. However some trial and error is involved when it is used to investigate the existence of a Γ_{H^2}-form. We will therefore only use it in deriving the subrearrangement Γ_L-criterion. This criterion is the desired solution to our problem. It is very easy to apply because it is systematic and only contains a few (μ) application of a single operation. ($\mu = c - r + 1$, the number of loops of Γ_L, is also called the cyclomatic number.)

Let us state the criterion:

The subrearrangement Γ_L-criterion. – A necessary and sufficient condition that a loop generator set Γ_L — in an explicitly independent form — has a pair Γ_H of Γ_{H^2}-form, is that a connected subgraph which corresponds to an arbitrary selection of loops (rows) of Γ_L can be i) — and iii-γ) — rearranged (see Lemma 2), so that the branches of the subgraph which are contained in another loop of Γ_L form a single path. The connected subgraph which also contains this new loop is obtained by connecting to the two end-vertices of the path, the other branches of the new loop, connected in series. If no branch of the loop is contained in the previous subgraph, the new loop is connected to the subgraph at one arbitrary vertex. If the process is started with a single loop of Γ_L as subgraph, it ends up with a desired H_T^2-graph, or if a path-rearrangement with operations i) or iii-γ) is impossible, then no H_T^2-graph exists. The order in which the loops are taken is arbitrary. Also if more than one rearrangement is possible, it is without significance which one is chosen.

Let us for the *proof* consider a Γ_L, Γ_H-pair

$$(19)\qquad \Gamma_L = \begin{array}{c} \\ l_1 \\ \cdot \\ l_i \\ l_{i+1} \\ \cdot \\ \cdot \end{array} \left\| \begin{array}{cccccc|c|ccccc} & & & & & & & \multicolumn{5}{c}{\overbrace{\qquad\qquad\quad \underbrace{\;\;}_{s_{i+1}}}^{s_i}} \\ 1 & 0 & 0 & 0 & 0 & 0 & \boxed{} & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & \boxed{} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & \boxed{} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & \boxed{} & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & \multicolumn{6}{c}{\boxed{}} \\ 0 & 0 & 0 & 0 & 0 & 1 & \multicolumn{6}{c}{\boxed{}} \end{array} \right\| \begin{array}{l} \\ \left.\begin{array}{c} \\ \\ \\ \end{array}\right\} \Gamma'_{L.i} \\ \\ \\ \\ \end{array}$$



$$(20)\qquad \Gamma_H = \begin{array}{r} 1 \\ 2 \\ 3 \\ \cdot \\ \cdot \\ r-5 \\ r-4 \\ r-3 \\ r-2 \\ r-1 \end{array} \left\| \begin{array}{cccc|cc|cccccccccc} \boxed{} & \boxed{} & \boxed{} & \boxed{} & \boxed{} & \boxed{} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & & & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & & & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right\| \begin{array}{l} \\ \\ \\ \\ \\ r_{i,1}(0), \\ r_{i,2}(0), \\ r_{i,3}(0), \\ r_{i,4}(0),\ r_{i+1,1}(0) \\ r_{i,5}(0),\ r_{i+1,2}(0) \end{array}$$

Γ_L is transformed into an explicitly independent form, *i. e.* with μ different columns only containing one 1. The corresponding (compare (3)) pair Γ_H (20) is then immediately obtained in explicitly independent form. The rows of the right part of Γ_L are the columns to the left of Γ_H.

We will temporarily assume that it is possible to construct a loop-graph corresponding to some of the rows of Γ_L, say the i first (in (19) the three first). Let the $\mu - i$ columns of the independent part of Γ_L which have 1's in the remaining $\mu - i$ loops of Γ_L be the ν''-partition of the Γ_{H^2}-criterion. We denote it with ν_i'' and we use the index i also for the ν'-partition ν_i': There may be a number of columns of $(\Gamma_L')_{\nu_i'}$ which are empty, say the s_i last. Withdrawing them from ν_i' we have the $\underline{\nu_i'}$-partition as indicated in (19). The $(\Gamma_H)_{\underline{\nu_i'}}$-matrix consists of s_i zero-rows $r_{i,1}(0)$, $r_{i,2}(0)$, ..., and the other rows are explicitly independent. According to our assumption, $(\Gamma_H)_{\underline{\nu_i'}}$ can be transformed into Γ_{H^2}-form. There are two possibilities for the corresponding subgraph. Either the graph G_i of the independent rows is nonseparable (Fig. 1-*a*) or it is separable (Fig. 1-*b*). If for instance in the last case there are s_i cut-vertices and s_i zero-rows, the graph can be split iii-β) of Lemma 2) into s_i+1 disconnected pieces (Fig. 1-*c*). In order to diminish the number of types of rearrangements, we can always require that G_i shall consist of only one piece. Even if G_i is separable and consists of disconnected pieces we know that the final H_T^2-graph shall be connected, but we do not know *a priori* how the pieces shall be connected. But if we connect them arbitrarily iii-α) we can always change their connectivity in any desirable way only with the operation iii-γ) of Lemma 2.

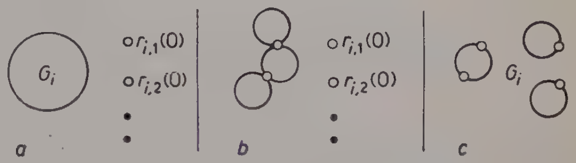


Fig. 1. – Possibilities for an H^2-graph.

Let us next consider the $(\Gamma_H)_{\underline{\nu_{i+1}'}}$-matrix, which corresponds to the same single loops as before and one further, say the $(i+1)$-th loop l_{i+1} of Γ_L. The $\underline{\nu_{i+1}}$-partition is indicated in (19) and (20). There are two kinds of possibilities.

1) The s_i-columns to the right in Γ_L are still empty in the first $i+1$ rows. Then $(\Gamma_H)_{\underline{\nu_{i+1}'}}$ has again s_i zero-rows, and the other rows are independent.

2) One or more $(s_i - s_{i+1})$ of the s_i-columns are non-empty in the first $i+1$ rows of Γ_L: The $(\Gamma_H)_{\underline{\nu_{i+1}'}}$-matrix has s_{i+1} zero-rows and the other rows are independent. (In (19) $s_i - s_{i+1} = 3$.)

Let us first consider case 1). We know that $(\Gamma_H)_{\underline{\nu_i'}}$ has a $(\Gamma_{H^2})_{\underline{\nu_i'}}$-form. $(\Gamma_H)_{\underline{\nu_{i+1}'}}$ only differs from $(\Gamma_{H^2})_{\underline{\nu_i'}}$ in that it contains one further column, the $(i+1)$-th. The content of this column is obtained by the product

$$(21) \qquad (\Gamma_{H^2})_{\underline{\nu_i'}} \cdot (\overline{l_{i+1}})_{\underline{\nu_i'}}$$

for we must have (compare (17))

$$(\Gamma_H)_{\underline{\nu'_{i+1}}} \cdot (\overline{l_{i+1}})_{\underline{\nu'_{i+1}}} = 0 \,. \tag{22}$$

(21) means that the $(i+1)$-th column of $(\Gamma_H)_{\underline{\nu'_{i+1}}}$ has a 1 in each row that corresponds to an end-vertex of a path-piece formed by the l_{i+1}-branches that are incident with the G_i-graph. It now follows from the Γ_{H^2}-criterion that a necessary and sufficient condition for $(\Gamma_H)_{\underline{\nu'_{i+1}}}$ to be of Γ_{H^2}-form is that it is possible to rearrange the connected G_i-graph with operations i) and iii-γ) so that the branches of it which are incident with the loop l_{i+1} form a single path (with two end-vertices). The G_{i+1}-graph is simply obtained by connecting the $(i+1)$-th branch between the two end-vertices of the path.

Let us then consider case 2) where $s_{i+1} \neq s_i$, *i.e.* $(\Gamma_H)_{\underline{\nu'_{i+1}}}$ contains $(s_i - s_{i+1})$ non-zero rows more than $(\Gamma_H)_{\underline{\nu'_i}}$ and the $\underline{\nu'_{i+1}}$-partition contains $(s_i - s_{i+1} + 1)$ branches more than the $\underline{\nu'_i}$-partition. (21) gives the content of the $(i+1)$-th column of $(\Gamma_H)_{\underline{\nu'_{i+1}}}$, here however only in the positions corresponding to the non-zero rows of $(\Gamma_H)_{\underline{\nu'_i}}$. Thus we see that if the path-pieces of the l_{i+1}-branches which are incident with the G_i-graph can be rearranged into a single path in G_i with operations i) and iii-γ) then we can transform $(\Gamma_H)_{\underline{\nu'_{i+1}}}$ into Γ_{H^2}-form in the following way. Row $r-4$ (see (20)) is added to row $r-3$. Next row $r-5$ is added to row $r-4$. Finally row $r-5$ is added to a row corresponding to an end-point of the path in the G_i-graph. The corresponding geometrical construction is that we connect to the two path end-points in G_i a path consisting of a series-connection of the remaining new loop-branches of l_{i+1}. The sufficient path-condition on G_i is also necessary for we know that all $(\Gamma_{H^2})_{\underline{\nu'_{i+1}}}$-matrices are generated with operations i) and iii-γ) of Lemma 2 from the one obtained. If the G_i-graph is separable it might be possible to rearrange the branches of the l_{i+1}-loop so that those branches which are incident with the G_i-graph form more than one path. But then the G_i-graph is not connected and it shall be according to the criterion when the G_{i+1}-graph is formed. After that, of course, any i) or iii-γ) operation on G_{i+1} is allowable. Finally if the intended path-branches of G_i cannot be rearranged into a single path, then no G_{i+1}-graph exists, nor an H_T^2-graph. For we know from the Γ_{H^2}-criterion that we can determine Γ_{H^2} by determining any $(\Gamma_{H^2})\underline{\nu'}$-partitions successively (compare (17)), and we have shown that the above $\underline{\nu'_i}$-partitions of an explicitly independent Γ_L-set are $\underline{\nu'}$-partitions. So if no single path of the l_{i+1}-branches can be made in G_i, the only construction would be to close the l_{i+1}-branches between the path-pieces in G_i (connected) so that the l_{i+1}-loop is formed. But then the number of new non-zero rows (real vertices) in (20) would be less than $s_i - s_{i+1}$, and hence this construction is impossible. This proves the Γ_L-criterion.

In the criterion the word « corresponding » is used in the same sense as was explained at the end of Sect. **3**.

The subrearrangement criterion is really a solution to the Okada-Seshu problem (*) (compare Sect. **1**), for as we shall see in the next section, there is no problem in rearranging intended loop-branches into a single path or in deciding whether a rearrangement is impossible.

5. – Path-rearrangements.

We will use the following expressions.

Separable subgraph: a subgraph which is connected to an internally connected subgraph at only one vertex, a cut-vertex [8].

Loosely tied, separable subgraph: a subgraph which is connected to an internally connected subgraph at precisely two vertices, and which contains at least one « internal cut-vertex » (a cut-vertex if the subgraph is regarded alone).

Loosely tied, nonseparable subgraph: the same as in the previous case but with no « internal cut-vertex ».

We want to decide if a number of intended loop-branches, say the path-branches p_i, can be rearranged with operations i) and iii-γ) of Lemma 2 so that a single path is obtained. The simple method to be described is quite straightforward. The procedure can be divided into a couple of independent steps.

Let us first consider the case when the graph contains a number of separable subgraphs containing p_i-branches. Since such subgraphs can only be connected at cut-vertices (*i.e.* so that no new loop is formed) it follows that the p_i-branches in a separable subgraph must be rearranged into a single path, however with no conditions on the end-points, since a separable subgraph can be reconnected iii-γ) at any vertex. The subgraphs are then reconnected so that a path of the p_i-branches is formed (see Fig. 2 *a* and *b*). The path is connected to an end-point of an eventual p_i-path in the main sub-

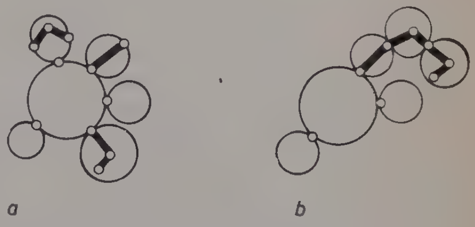


Fig. 2. – Rearrangement of separable subgraphs for a path-formation.

(*) Recently R. Gould [1] has given a systematic treatment of the problem, however far more complicated than the solution given here.

graph. For simplicity one could omit separable subgraphs which do not contain p_i-branches. We have now referred the investigation to a non-separable graph.

We consider a loosely tied, non-separable subgraph g_i, only connected to the remaining graph at the two change-around vertices v_{i1} and v_{i2}. Even if g_i and the remaining graph contain other change-around vertices, no branch which from the beginning is not contained in g_i can be rearranged into g_i. So if g_i does not contain a p_i-branch, it could for simplicity be replaced by a single branch. If it does contain p_i-branches, these can only be brought into contact with p_i-branches outside g_i at the two vertices v_{i1} and v_{i2}. Hence the p_i-branches in g_i must be rearranged so that they form paths in g_i with end-points at v_{i1} or v_{i2}. We will here distinguish between two cases:

1) All p_i-branches outside g_i form a single path with its two end-points at v_{i1} and v_{i2} respectively.

2) All other cases concerning the p_i-branches outside g_i.

The distinction between the two cases is immediately recognized. If a p_i-path exists outside g_i so that it is incident with both v_{i1} and v_{i2} then the path-piece between v_{i1} and v_{i2} always exists (it cannot be broken by a rearrangement). So if there are no more p_i-branches outside g_i than the mentioned path-piece we have case 1), otherwise case 2) for in the second case not both v_{i1} and v_{i2} can be end-vertices. Also, of course, in case 2) we do not need to have a path between v_{i1} and v_{i2} outside g_i.

In case 1) there are two possibilities for the p_i-branches inside g_i. They must either be rearranged into a single path with one end-point at v_{i1} or at v_{i2}. Or they must be rearranged into two single paths, one with an end-point at v_{i1} and the other with an end-point at v_{i2}.

In case 2) there is only one possibility: the p_i-branches inside g_i must be rearranged into a single path with one end-point at v_{i1} or at v_{i2} (or with both end-points at v_{i1} and v_{i2} respectively).

If these possibilities for the two cases are not at hand then a resulting p_i-path cannot be formed.

If however a loosely tied, non-separable subgraph does not exist, the whole graph must consist of a single loop (after the indicated simplifications). Any pair of non-adjacent vertices is then a change-around pair but the corresponding loosely tied subgraphs are separable for they contain internal cut-vertices. In this case it is clearly always possible to rearrange the p_i-branches so that they form a single path. For simplicity we could reduce the number of non-p_i-branches in a pure series-connection of branches to one.

The whole procedure is now simply this. Start with a loosely tied, non-separable subgraph g_i containing at least one p_i-branch and not containing a smaller p_i-containing loosely tied, non-separable subgraph. If case 1), make the above-mentioned rearrangements. I they cannot be done with a pure series-interchange of the branch-order no resulting single path exists. If case 2), make the above-mentioned corresponding rearrangement and simplifications. If the rearrangement cannot be done with a pure series-interchange of the branch-order, no resulting single path exists. If the rearrangement is possible the loosely tied, non-separable subgraph should be enlarged to another subgraph of the same kind and the procedure is repeated. An enlargement is only possible if there exists a change-around vertex pair $v_{i+1,1}$, $v_{i+1,2}$ so that one of these vertices is incident with one vertex of the previous pair $v_{i,1}$, $v_{i,2}$ and so that the other v_{i+1}-vertex is not contained in the g_i-graph. The process ends up with a desired resulting p_i-path or with a decision that no such path exists. It is of course also possible to start from different subgraphs, each being successively enlarged.

Corollaries from this general treatment are the following tests.

If none of the vertices of a not properly placed p_i-branch (end-points of a path-piece) is incident with a vertex of a change-around pair, then no path can exist (see Fig. 3-*a*).

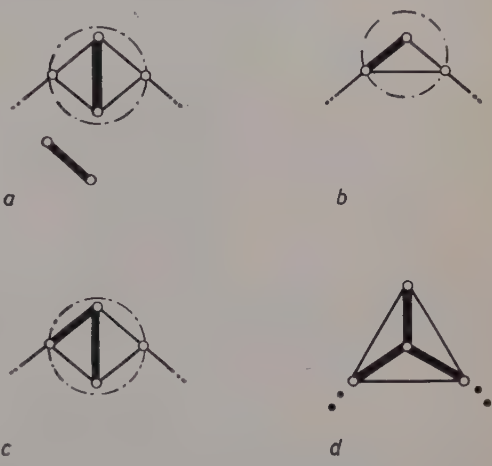


Fig. 3. – Examples of p_i-configurations for path-tests.

If only one of the vertices of a not properly placed p_i-branch (end-points of a path-piece) is incident with a vertex of a change-around pair (see Fig. 3-*b* and 3-*c*), and if more than two such branches (path-pieces) exist, then no resulting path can be formed.

If more than two p_i-branches are completely tied at a single vertex, then no path can exist (Fig. 3-*d*).

6. – Examples.

In the first example given here we will apply the *c*-criterion and the Γ_L-criterion to a Boolean function in order to investigate if it has an irredundant network. In the second example we want to decide if a Γ_H-set can be transformed into Γ_{H^2}-form. For this we apply the Γ_L-criterion to a Γ_L-pair of Γ_H. The third example, finally, will illustrate that it is in general necessary that

Γ_L is first transformed into an explicitly independent form before the Γ_L-criterion can be applied.

Example 1. – We want to investigate if

$$B = ab \cup dfg \cup ac'deg \cup cefg \cup bc'ef$$

has an irredundant network. B contains no redundant literals, for it contains only prime implicants (compare [3]). The corresponding L_T-matrix is

$$\Gamma_T = \begin{array}{||ccccccccc||}
T & a & b & c & c' & d & e & f & g \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\
1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0
\end{array}$$

L_T is transformed with row-replacements into an explicitly independent Γ_L-form:

$$\Gamma_L = \begin{array}{||ccccccccc||c}
T & a & b & c & c' & d & e & f & g & \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\
0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 2 \\
0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 3 \\
0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 4
\end{array}$$

Γ_L generates the coset c:

$$c = \begin{array}{||ccccccccc||}
T & a & b & c & c' & d & e & f & g \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\
1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0
\end{array}$$

and hence we obtain for $B(c)$:

$$B(c) = B \cup abcd \cup bcc'df \cup acc'g \, .$$

The c-criterion is fulfilled, for the two last intersections vanish, and the other indicated clause of $B(c)$ subsumes the clause ab of B.

Let ut now apply the Γ_L-criterion for instance in the loop-order 4, 3, 2, 1. The four steps are indicated in Fig. 4. First loop 4 of Γ_L is drawn (Fig. 4-a). The p_i-branches for loop 3, e and d, are already in a single path, and the corresponding graph is drawn (Fig. 4-b). In order to form a path of the p_i-branches c' and e for loop 2 the series-connected branches g and c' are interchanged. The graph also containing loop 2 is drawn (Fig. 4-c) and the p_i-branches for the final loop 1 are indicated. We have here three p_i-branches and two of them, for instance d and g are contained in a loosely tied, non-separable subgraph (indicated in Fig. 4-c). We have here case 2) because no path of p_i-branches outside the subgraph between its connection vertices exists. Thus the two branches d and g must be rearranged into a path in the sugbgraph with an end-point at one of its connection vertices. So the series-connected branches b and g are interchanged. After that the subgraph is turned around and the desired path is formed. The remaining loop can now be adjoined and the irredundant graph of B is shown in Fig. 4-d.

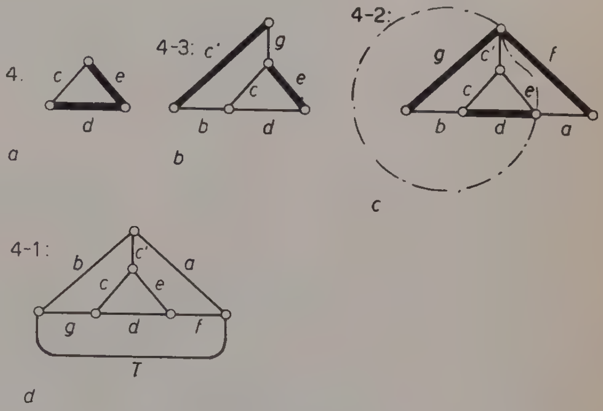


Fig. 4. – G_i-constructions for example 1.

Example 2. Let us investigate if the following Γ_H-set

$$\Gamma_H = \begin{array}{c} \begin{array}{cccccccc} a & b & c & d & e & f & g & h \end{array} \\ \left\| \begin{array}{cccccccc} 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right\| \end{array}$$

$$\Gamma_L = \left\| \begin{array}{cccccccc} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \end{array} \right\| \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array}$$

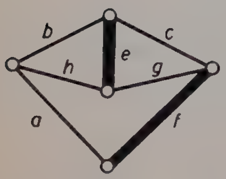


Fig. 5. – G_i-graph for example 2.

can be transformed into Γ_{B^2}-form, *i.e.* let us investigate if the indicated Γ_L-pair passes the Γ_L-criterion or not. The first steps corresponding to the three first loops of Γ_L give the G_i-graph of Fig. 5 where the p_i-branches for loop 4 are indicated.

The branch e is not incident with a change-around vertex and since the two p_i-branches e and f do not already form a path, we conclude that no Γ_{B^2}-pair to Γ_L exists.

Example 3. – Let us consider the loop generator-set Γ_L:

$$\begin{array}{cc} & \begin{array}{cccccccc} a & b & c & d & e & f & g & h \end{array} \\ \Gamma_L = & \left\| \begin{array}{cccccccc} 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \end{array} \right\| \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} \end{array}$$

A corresponding explicitly independent set Γ'_L is:

$$\Gamma'_L = \left\| \begin{array}{cccccccc} 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{array} \right\| .$$

According to the Γ_L-criterion we must apply it to an explicitly independent form. We want to see however what can happen when the criterion is applied on the above Γ_L-form. After the three first steps (the three first loops of Γ_L) we obtain the G_i-graph of Fig. 6-*a*. The process cannot be carried out further

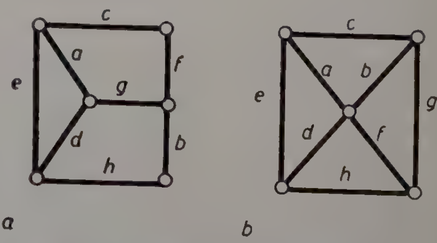


Fig. 6. – Graphs of example 3.

because the fourth loop does not contain any new branches. Although all the loops of G_i are correct (compare the complete graph of Fig. 6-*b* which is obtained when the criterion is correctly applied on Γ'_L), the structure of G_i has nothing to do with the structure of the complete graph.

We want to point out also that it can happen that there is a new branch at each step of a Γ_L-form which is not explicitly independent. Anyhow the Γ_L-criterion should not be applied until Γ_L has been transformed into explicitly independent form for it can otherwise happen that a row of Γ_L represents a multiple loop.

* * *

This work has been supported by the Swedish Research Institute of National Defense and by the publishing-house Natur och Kultur, Stockholm.

REFERENCES

[1] R. Gould: *Journ. Math. and Phys.*, **37**, 193 (1958).
[2] S. Okada: *Proc. Symp. on Information Networks*, Polytechnic Institute of Brooklyn, (1954).
[3] W. V. Quine: *Amer. Math. Month.*, **62**, 627 (1955).
[4] S. Seshu: *IRE Trans.*, Vol. CT-3, 172 (1956).
[5] C. E. Shannon: *Trans. Am. Inst. Electr. Eng.*, **57**, 713 (1938).
[6] C. E. Shannon: *Bell Syst. Techn. Journ.*, **28**, 59 (1949).
[7] O. Veblen: *Amer. Math. Soc.*, Colloquium Publications. Part II, vol. V (1931).
[8] H. Whitney: *Amer. Math. Soc.*, **34**, 339 (1932).
[9] H. Whitney: *Amer. Journ. of Math.*, **55**, 245 (1933).

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3º Trimestre

Analysis of a Non-Linear Biological Filter.

W. REICHARDT

Forschungsgruppe Kybernetik, Max Planck Institut für Biologie - Tübingen.

The stimulus reaction systems which have been studied most intensively by physiologists are threshold systems, or so called all-or-one systems. They give no graded response to a stimulus input signal. The analysis of such systems is therefore a very difficult one.

The situation is obviously much simpler when the system under consideration answers in a graded manner to stimuli. Here one can be more hopeful to carry the analysis of the transforming filter down to the differential equations involved.

Such a system which responds in a graded manner to light stimuli is the growth organ of the sporangiophores of *Phycomyces*. It serves as a receptor to stimuli, evaluates the absorbed information and serves finally as an effector system.

The sporangiophores of *Phycomyces* are parts of single cell systems. Each of them carries a sporangium on its upper end. Growth is confined to a zone 3 mm in length immediatly below the sporangium. In the growing zone the cell wall stretches and new wall material is built into the « gaps ». The grownig zone stretches without getting longer. For every amount that it grows in length a corresponding amount at its bottom is converted into secondary cell wall which has ceased to grow in length. In this stage of growth—after formation of sporangium—a steady growth rate is maintained for many hours (ERRERA, 1884 and CASTLE, 1942). The rate amounts about 3 mm per hour. The stretch along the growing zone is not distributed in a homogeneous manner. The strength of stretch is maximum 0.5 mm below the sporangium, stays on a nearly constant plateau between 0.7 and 1.9 mm and finally drops down (COHEN and DELBRUECK, 1958).

The sporangiophores are positively phototropic. When they are exposed to light from one side they grow towards the light. BLAAUW (1914) discovered another effect of light on the growth of the sporangiophores: the light growth response. This response refers to the situation in which the specimen is at all times symmetrically illuminated from two or more sides. If the illumination is symmetric with respect to the vertical axis and if the specimen is growing vertically at the start of the experiment this illumination will not cause it to deviate from vertical growth. If the illumination is kept at a constant intensity for a certain length of time the rate of growth is also constant and is the same whatever the intensity. However a striking and transient change in growth rate occurs when the intensity of illumination is changed.

We shall deal first of all with this effect of the so called light growth responses. In this case the input variable is $I(t)$, the light intensity as a function of time. The output variable, we measure, is the growth velocity $v(t)$ of a sporangiophore. The connecting process which transforms $I(t)$ into $v(t)$ is the unknown filter (DELBRUECK and REICHARDT [1]).

In order to analyse the filter process we submit the growing zone to different stimuli, measure the growth reaction and draw conclusions from these functional data on the filter process involved. The first mentioned experimental fact is

$$(1) \qquad v = v_0 = \text{const} \qquad \text{if} \qquad I(t) = \text{const}\,.$$

The rate of growth is constant if the intensity of light is kept constant whatever the amount of intensity. In the second step we raise the question whether the transformation between $I(t)$ and $v(t)$ is a linear one or not. Let us call the first test stimulus—which refers to a special change of intensity—$I_1(t)$ and the connected growth velocity deviation $Dv_1(t)$. The second stimulus (submitted to the growing zone after the reaction to the first one had died out) we call $I_2(t)$ and the connected growth velocity deviation $Dv_2(t)$. If the transformation between $I(t)$ and $v(t)$ would be a linear one, we should find that a superposition of the two stimuli should be followed by a reaction being the superposition of the single reactions to the programs I_1 and I_2. The experiments showed to be in flat contradiction to this; the superposition rule does not hold. It means the filter process between $I(t)$ and $v(t)$ is obviously a non-linear one.

We consider now a special light program. The plant was illuminated with constant intensity I_1 for a while (for inst. 30 min). At time $t = t_0$ we superimposed a light flash during a time Δt (for inst. 15 s). More exactly the intensity I_1 jumped to the intensity I_2, was kept constant for the time Δt and finally switched down to the original intensity I_1. The growth velocity reaction to such a light program showed the following behaviour: after the stimulus growth continues at its normal rate for 2.5 minutes, then increases for a few minutes to a maximum which may be twice as high as the normal rate. Presently it decreases again, goes below normal, and returns to normal by about

15 minutes after the stimulus. The net gain in growth due to such a stimulus is zero, *i.e.* the transient increase in growth rate is compensated for by the subsequent fall below the normal level. The stimulus does not produce extra growth, it simply alters the distribution in time of the growth that would have taken place during the same period in the absence of the stimulus.

If one repeats the last experiment, keeping the adapting intensity I_1 constant, but varying the intensity or duration of the stimulus, we find that the reaction is a graded one. Secondly we find that a change in the stimulus does not alter the latent period. Thirdly we find that the shape of the response curve is independent of the stimulus except for very large stimuli. Fourth we find that the responses depend on the product intensity × time as long as the duration of the stimulus does not exceed the range of one minute.

From what we have just said about the general characteristics of the response as a function of stimulus, it is clear that it is not necessary to measure the entire response curve each time in order to get a measure of the intensity of the response. We have chosen as a measure of the response the ratio of the growth during the period from 2.5 to 5 minutes after the stimulus to that during the period from 0 to 2.5 minutes. The period from 2.5 to 5 minutes takes in most of the positive phase of the response, and the period from 0 to 2.5 minutes gives the base line of normal growth. This ratio will be called R.

If one considers the adapting intensity I_1 as a parameter and varies the stimulus $S = I_2 \cdot \Delta t$, R turns out to be a logarithmic function of S/I_1,

$$(2) \qquad R \sim \log S/I_1 \, .$$

In other words, the stimuli have to be increased proportionally to the adapting intensities in order to produce the same growth velocity reactions. This relation is analogous to the well known Weber-Fechner law. These experimental findings give information about the sensitivity of the speciments after they had been brought to equilibrium with an adapting intensity of illumination.

We wish now to introduce a measure for the level of adaptation *i.e.* we want to introduce a quantity which in some manner characterizes the sensitivity of the specimen at any given moment, whether it is in equilibrium or not. The measure of adaptation will be most accurate if it utilizes responses in a region where the response changes most rapidly with the stimulus. A response $R = 1.4$, which is produced by $S/I_1 = 2^3$ minutes is a suitable choice for this purpose. Our measure of the level of adaptation should obviously be proportional to the critical stimulus. One could make the proportionality factor equal to one, but this would be somewhat arbitrary since the critical stimulus was defined with the aid of an arbitrarily chosen standard response. A more rational choice is achieved by the following line of thought. We ask: with which intensity would we have to equilibrate the specimen in order to bring it to the same level of adaptation? This we will call the equivalent intensity. It is obtained by dividing the critical stimulus by 8 minutes. We will give this intensity the name A. After equilibration with I_1, the level of adaptation is by definition $A = I_1$. One is now in a position to outline a procedure for determining A also for some non-equilibrium states. The procedure consists in bringing the speciment into the particulare state, testing it with various stimuli, determining by interpolation the stimulus giving response $R = 1.4$ and dividing this stimulus by 8.

Before proceeding with the presentation of our experimental data we should briefly restate the immediate goal. Our observational data concern with growth velocities in response to certain illumination programs. We speak of the growth output in response to an illumination input. The functional relation between these two is enormously influenced by a variable describing the internal state of the specimen, which we have called the level of adaptation A. We know that the variable itself is determined by the illumination program, and it constitutes therefore another and perhaps more immediate output of the illumination input. To explore this relation we measure A as a function of time after equilibration with various intensities or to a superimposed short stimulus of various sizes.

The experimental results reveal the following facts about the connection between I and A. First of all we find that in each case of illumination A drops down exponentially by a factor two in 2.5 minutes after the light is turned off. Secondly we find in the case of a superimposed light flash of strength S that the level of adaptation rises from the level I_1 to the level

$$A = I_1 + S/b\,, \tag{3}$$

where I_1 is the level of adaptation to which the plant had been adapted before and $b = 3.8$ minutes (e-time) the time constant of the adaptation system. From these experimental findings we draw the conclusions, that the functional relation between $I(t)$ and $A(t)$ is described by the differential equation

$$b\,(\mathrm{d}A/\mathrm{d}t) + A = I\,. \tag{4}$$

This equation obviously satisfies the basic findings that $I = A$ after equilibration with constant intensity, that A decreases in the dark exponentially with the time constant b, irrespective of its initial level and irrespective how it was brought to this level. For an arbitrary illumination program $I(t)$ the equation has the integral

$$A(t) = A(0)\exp\left[-t/b\right] + (1/b)\int_0^t I(s)\exp\left[\frac{-(t-s)}{b}\right]\mathrm{d}s\,. \tag{5}$$

The last equation implies that during a short stimulus A increases by $(1/b)\int I(s)\,\mathrm{d}s = S/b$ in accordance with the last mentioned experiment.

In the next step we have to deal with the coupling between illumination input, adaptation and growth output. The experimental results have shown that stimuli and adaptive levels have to be raised proportionally to produce the same growth output. We infer that what is relevant for the growth output is the ratio I/A. This functional parameter we call $i(t)$.

Let us consider now what happens to $i(t)$ during and after a short stimulus S superimposed upon a constant background intensity I_1. During a short square shaped stimulus of intensity I_2 and duration Δt, A increased linearly from $A_- = I_1$ to $A_+ = A_- + S/b$. Therefore during the stimulus we have the relation

$$A(t) = A_- + St/b\Delta t\,. \tag{6}$$

Since this increase of A during the stimulus may represent an increase by a large factor, $i(t)$ jumping from unity to a high value at the beginning of a stimulus, may decrease by a large factor even during the shortest stimulus. This decrease has to be taken into account in evaluation of the spike transient in the functional parameter $i(t)$, which we conceive to be the quantity immediately responsible for the growth output. Let us define an internal stimulus s of the system as the integral of the transient during the stimulus. It works out as follows

$$s = \int_0^{\Delta t} (I_2/A)\,\mathrm{d}t = I_2/(A_- + St/b\,\Delta t)\,\mathrm{d}t = b\log\,(1 + S/bA_-)\,. \tag{7}$$

For strong stimuli, *i.e.* for stimuli which are in the range of those which give the standard response or stronger, s ist a logarithmic function of S in agreement with the above mentioned experiments. For small stimuli (small to those which give the standard response) the logarithm may be developed into a power series of S/bA_- and s becomes equal to the first term of the series. We then have $s = S/A_-$, s is proportional to S. An experimental test of this last prediction seems impossible hence the fluctuation of the growth velocity does not permit to test this range of the reaction.

The functional parameter $i(t)$ has another important property. For an arbitrary illumination program which is preceded and followed by equilibration with the same intensity I_1 we have

$$\int (i-1)\,\mathrm{d}t = \int \left(\frac{I-A}{A}\right)\mathrm{d}t = \int_{I_1}^{I_1} b\,\frac{\mathrm{d}A}{A} = 0\,. \tag{8}$$

The relation means that the positive and negative deviations of i resulting from any program which returns to the original intensity cancel exactly. This prediction was checked very carefully with two light programs, a pulse up and a pulse down program superimposed on a constant adapting intensity I_1. The results show very clearly that there is not net gain or loss in growth in both of these experiments. If the program goes from equilibration with I_1 through any intermediate course to equilibration with another intensity I_2, then we have the relation

$$\int (i-1)\,\mathrm{dt} = b \log (I_2/I_1)\,. \tag{9}$$

The last relation was checked by step up and step down light programs and found in agreement with the above mentioned statement.

These relations suggest that the functional parameter $i(t)$ may be the important variable which quite generally stands in a linear functional relationship to the growth output. By this we mean the following: At equilibrium $i(t)$ rs always unity. Under the influence of a given illumination program it will deviate from unity in a predictable manner. Let us assume we could design an illumination program resulting in a $i(t)$ program which deviates from unity only during a short period. Such a pulse in $i(t)$ will lead to a growth output iepresented by a certain function of time. We now postulate two things: First that the growth output for pulses in $i(t)$ of various sizes is equal to the output produced by a unit pulse in $i(t)$ multiplied by the actual size of the pulse and secondly, that the growth output for an arbitrary illumination program can be calculated as a simple superposition of the outputs of all the pulses into which the functional parameter $i(t)$ can be decomposed. These two postulates are formulated analytically in the next equation, in which $Dv_1(t)$ represents the growth output due to a unit pulse in $i(t)$, $Dv(t)$ represents the actual output resulting from an arbitrary illumination program, and $Di(t)$ represents the $i(t)$ output of the program. D expresses that we are referring to deviations from the equilibrium values of the velocity and of $i(t)$, respectively

$$Dv(t) = \int_{-\infty}^{t} Dv_1(t-s)\,Di(s)\,\mathrm{d}s\,. \tag{10}$$

We have attempted to test the linear functional relationship between $i(t)$ and $v(t)$ in another manner which is less direct but more accurate, involving medium size periodic stimulations superimposed upon a constant intensity and comparing the growth outputs for two such programs in which the periods differ by a factor two. We cannot predict the v output for either one of these programs until we know the basic response function $v_1(t)$. However it can be shown very easily that the v outputs of the two programs should be related to each other by the following equation

$$Dv_T(t) = Dv_{2T}(t) + Dv_{2T}(t+T)\,, \tag{11}$$

where T is the shorter of the two periods. This equation may be expressed by saying that the v output of the T program is obtained from the v output of the $2T$ program by superimposing the first and second half of the latter over each other. The predicted curves show very good agreement with the experimental curves. Actually the application of equation (11) presupposes that the specimens come to adaptive equilibrium with intensity I_1 in the intervals between stimulations. In our experiments this condition was met well in the $2T$ program but somewhat imperfectly in the T program. As a result, the stimuli in $i(t)$ were probably a little smaller in the T program than in the $2T$ program. The v outputs should be proportional to these stimuli. The size of this correction depends on the precise value of the time constant b of dark adaptation experiments, the correction amounts to 25%. This seems rather more than our experimental error.

All these experimental findings can be concluded in the functional structure given below

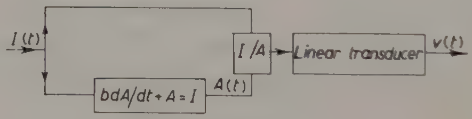


Fig. 1.

The light $I(t)$ controls the growth velocity $v(t)$ in two ways.

1) $I(t)$ determines the adaptive level $A(t)$ with the time constant b of 3.8 minutes. The quantitative relationship between I and A is described by equation (4).

2) $I(t)$ determines the functional parameter $i(t)$. The deviation $Di(t)$ controls $Dv(t)$, the deviation of growth by a linear transducer. Since the analysed filter connecting $I(t)$ with $v(t)$ only contains differential equations and one basic logical operation it transforms the « field of stimuli » into the « field of reactions ».

Up to this point we have studied reactions to stimulations of the entire growing zone. Such an experimental procedure does not give any information about the question whether the functional structure of the filter process is built up by complicated interaction phenomena or simply by superposition of light growth reactions carried out by the parts (molecular groups) of the growing zone themselves. Let us take the last mentioned possibility as an hypothetical statement and ask: How could such a functional autonomy of molecular groups in the growing zone be tested? The answer can be split into two parts; first the test for azimuthal and second the test for longitudinal autonomy of growth reactions (with respect to the geometry of the sporangiophore).

The first test was carried out during the last months (REICHARDT and VARJU [2]), the second one is still incomplete. The idea is as follows: When we studied the growth reactions, the specimens were illuminated bilaterally in such a way that the light distribution was a symmetrical one. The situation is quite different when the growing zone receives light only from one side. In fact the zone acts as a converging cylindric lens. It concentrates the entering light on an area of only 20 % of the backward side of the growing zone. Therefore we deal with an asymmetrical-light distribution. The result is the well known positive phototropic reaction. The sporangiophore bends with an angular velocity which remains constant as long as the light beam falls perpendicularly onto the growing segment (above the bend). The absolute amount of the angular velocity is independent of the light intensity within the large range from 0.1 to 200 erg/cm² s (the «normal range») and drops down for higher and lower intensities The phototropic reaction disappears for $3 \cdot 10^{-6}$ erg/cm² s on the lower end and for $2.2 \cdot 10^{3}$ erg/cm² s on the upper end of the intensity range. These findings in mind, we designed the following experiment in order to study the azimuthal autonomy. The specimen were bilaterally adapted for nearly 50 minutes to an intensity I_1 near the lower end of the normal intensity range. For the time $t = t_0$ one of the light channels was switched off and the other kept at the same intensity till the upper end of the sporangiophore reached the « normal » stationary angular velocity. Now the intensity I_1 of the light channel is suddenly raised to an intensity I_2 (also within the normal intensity range) and kept constant. What can be predicted on the phototropic reaction on such a light program? If there exists azimuthal autonomy with respect to growth reactions the net gain in growth on the fully illuminated side of the growing zone should be about five times larger than the net gain of the other cell side. This is partly a consequence of equation (9). We have to expect an inversion phase of the phototropic reaction as an answer to the one channel step function program. The experiments showed the predicted effect very clearly. After the specimen had received the light step function their positive phototropic reaction turned into a negative one. After this phase it came back to the normal stationary behaviour of positive phototropism when the autonomous parts of the growing zone have adapted to the new intensity I_2. Before and after the inversion phase the angular velocity of the sporangiophore is the same since in both cases we stay in the normal phototropic range. The net effect of the reaction consists in a time shift Δt of the phototropic reaction curve. Δt turns out to be a logarithmic function of the intensity ratio I_2/I_1. If one connects mathematically the time shift Δt with the integral stretch of the growing zone and determines the integral stretch to the step function program, the theory yields the relation

$$\overline{\Delta t} \sim \log I_2/I_1 \,, \tag{12}$$

which is in accordance with the experiments. These findings favour the assumption of autonomous parts of the growing zone.

Finally we have to raise the question, is it possible to explain the stationary phototropic response by the functional mechanism of the light growth reaction. We should like to propose the following model: BUDER (1918, 1920) has converted the converging lens property of the growing zone into a diverging lens property by immersing the sporangiophore in a medium of higher refractive index than that of the protoplasm of the sporangiophore. Under these conditions the positive phototropism was converted into a negative one. In addition our experiments have shown that in the normal range of phototropic response the angular velocity of the sporangiophore is independent of the illumination intensity. From here we draw the conclusion that the intensity distribution of light on the surface of the growing zone determines the phototropic reaction. If one takes in account the circulation of the cell wall material around the axis of the sporangiophore one has to consider the growth output effects of the autonomously reacting parts as a function of the light intensity distribution on their way. Rough calculations have shown that such a model would fulfil the main finding: the positive phototropic reaction to an illumination from one side. The experiments on this part of the analysis are still incomplete.

REFERENCES

[1] M. DELBRUECK and W. REICHARDT: *System analysis for the light growth reactions of phycomyces. Cellular mechanism in differentiation and growth. Fourteenth growth Symposium* (Princeton, 1956).

[2] W. REICHARDT and D. VARJU: *Zeits. f. Phys. Chem.*, **15**, 267 (1958).

SUPPLEMENTO AL VOLUME XIII, SERIE X
DEL NUOVO CIMENTO

N. 2, 1959
3º Trimestre

A Cross Correlation Process in the Nervous Center of an Insect Eye.

B. Hassenstein

Forschungsgruppe Kybernetik Max-Planck-Institut für Biologie - Tübingen

Vision of movement in all animals and men involves a physiological interaction between adjacent visual units. In the beetle *Chlorophanus viridis* this interaction was shown to be a process of cross-correlation. The eyes of this insect are composed of facets (ommatidia) which act as visual units in the process of perception of movement. The visual fields of adjacent ommatidia do not overlap. One point-like visual stimulus is received only by one ommatidium and not by its neighbors. The anatomical angle between the axis of two adjacent facets is 6.8°. Many animals including *Chlorophanus* react to the perception of movement in their visual field by optomotor reactions. They follow the movement which they perceive by active turning reactions of their head or their body and so reduce the movement stimulus which they receive by their eyes. This may be described in terms of a feed back loop.

The direction and strength of the optomotor response has been used as an indicator of the perception processes in the nervous parts of the eyes of the experimental animal. In the experiments the feed back loop of the reaction has been cut-off by fixing the animal so that its optomotor reactions could be observed by the experimenter but did not influence the position of the animal itself in relation to its optical environment. The experimental procedure (Y-maze-globe method) has been described (1951, 1958) in full detail elsewhere.

Successions of practically point-like light stimuli were delivered to the eye of the experimental animal. If A, B, C, D, ... are adjacent ommatidia in an horizontal row and if the sign « + » is given to a stimulus which consists of an illumination change from darker to lighter the formula ${}^{+}A(t_1)\,{}^{+}B(t_2)$ may describe a succession of two stimuli in adjacent ommatidia. The same succession may be written also $F_{AB}^{++}(t_1, t_2)$. The reaction of the animal to F_{AB}^{++} may be symbolized by R_{AB}^{++}.

1. - Results.

1) The simplest succession of light changes which is able to release an optomotor response consists of *two* stimuli in adjacent ommatidia.

2) In producing optomotor responses each ommatidium can only cooperate with its immediate neighbor or with the next but one. There is no physiological interaction between ommatidia which are separated by more than one unstimulated ommatidium.

3) The maximum reaction is given in the case of a time interval between two stimuli of about 150 ms. The strength of reaction decreases with both greater and smaller time intervals. The maximum time interval which was shown experimentally to release a reaction was slightly over 10 s. One must conclude that the first stimulus has an after-effect of 10 s or more which later disappears. The real physiological interaction takes place between the after-effect of one stimulus and the effect of a following one. The first stimulus of a succession of two stimuli is modified by a filter which acts like a low pass.

4) $R_{BA}^{++} = -R_{AB}^{++}$.

5) $R_{AB}^{--} = +R_{AB}^{++}$.

6) $R_{ABC}^{+++} = R_{AB}^{++} + R_{BC}^{++} + R_{AC}^{++}$.

7) $R_{AB}^{+-} = R_{AB}^{-+} = -R_{AB}^{++} = -R_{AB}^{-}$

i.e. F_{AB}^{+-} releases a *negative* optomotor reaction.

8) As it has been known for a long time, a cylinder of gray stripes on white background releases weaker optomotor reactions than a cylinder of black stripes on white background which rotates with the same angular velocity. The strength of optomotor reactions of insects does not only depend on the velocity of the moving pattern but also on the amount of stimulus efficiency of the individual light changes of which the stimulus situation consists. In the following experiment I kept the time intervals of the stimulus succession constant and varied the stimulus intensities by using patterns of different gray shades with different contrasts.

The result of this experiment was: The strength of reaction is a quadratic function of the stimulus intensities.

The experimental results 7) and 8) may be described together as follows: The direction and intensity of the optomotor response reflect the multiplication result of signs and intensities of the individual stimuli. There must be a physiological mechanism which causes that the sensory input and the motor output are linked by a process which works according to the formula of multiplication.

The experimental facts 1)-8) may be represented by Fig. 1. The omma-

INDICE DEL SUPPLEMENTO

AL VOLUME XIII, SERIE X, DEL

NUOVO CIMENTO

Anno 1959

Fine del *Supplemento* al Vol. XIII, Serie X
del *Nuovo Cimento*, 1959

Direttore responsabile: G. POLVANI Tipografia Compositori - Bologna

Questo fascicolo del *Supplemento* è stato licenziato dai torchi il 14-XII-1959

1432

tidia may be represented by A and B. The low pass filters are symbolized by their time constant ^{+}E and the multiplication process by M.

9) W. Reichardt calculated the output of the system of Fig. 1 for the following conditions:

a) Input A and B are stimulated by a moving pattern which is at random composed of shades from white through black in such a way that both A and B receive « white noise » stimulation.

b) The output of the channel of B is subtracted from the output of the channel of A (since both of them represent movements of opposite directions).

c) The transmission is supposed to be linear in both the two low pass filters and the two other filters which represent the inertia of the conduction lines which cross between the two straight channels; the time constants of the filters are symbolized by τ_E and τ_D.

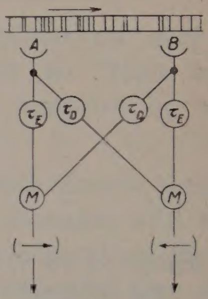


Fig. 1.

Reichardt found the output to be given by

$$f(v) = \text{const}\cdot\left(\exp\left[-\frac{x}{v\tau_E}\right] - \exp\left[-\frac{x}{v\tau_D}\right]\right).$$

Thereby v is the velocity of the stripe pattern relative to the sensory inputs. x is the spatial distance between the inputs.

This theoretical result has been tested in the beetle *Chlorophanus* by measuring the strength of optomotor response to the movement of a pattern which was at random composed of stripes of 10 shades of gray, each of the width $0.5x$. The result matched the theoretical curve very well if the time constants τ_E and τ_D were given the values 3500 ms and 46 ms.

2. - Conclusion.

The experimental facts suggest that in the eye of *Chlorophanus* the evaluation of movement in the visual field is made by a correlation process like that shown in Fig. 1.

BIBLIOGRAPHY

B. Hassenstein: *Zeits. f. vergl. Physiol.* **33** (1951) and **40** (1958).
B. Hassenstein and W. Reichardt: *Zeits. f. Naturforsch.*, **8** b (1953) and **11** b (1956).
W. Reichardt: *Zeits. f. Naturforsch.*, **12** b (1957).
B. Hassenstein: *Zeits. f. Naturforsch.*, **13** b (1958).